AGARD-CP-487

AGARD CONFERENCE PROCEEDINGS No.487

Bridging the Communication Gap

(Combler les Lacunes dans le Domaine de la Communication)

AGARD-CP-487

NORTH ATLANTIC TREATY ORGANIZATION

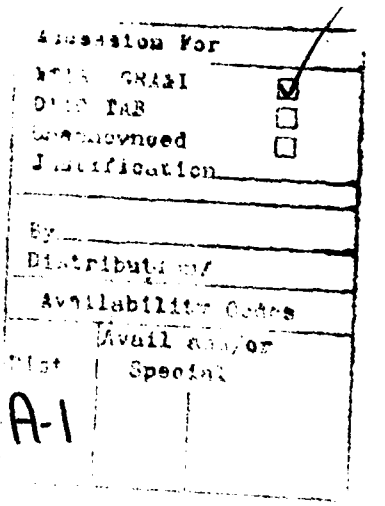
ADVISORY GROUP FOR AEROSPACE RESEARCH AND DEVELOPMENT

(ORGANISATION DU TRAITE DE L'ATLANTIQUE NORD)

AGARD Conference Proceedings No.487

Bridging the Communication Gap

(Combler les Lacunes dans le Domaine de la Communication)



Copies of papers presented at the Technical Information Panel Specialists' Meeting held at the Nye Sentrum Hotel, Trondheim, Norway, 5th to 6th September 1990.

The Mission of AGARD

According to its Charter, the mission of AGARD is to bring together the leading personalities of the NATO nations in the fields of science and technology relating to aerospace for the following purposes:

- Recommending effective ways for the member nations to use their research and development capabilities for the common benefit of the NATO community;
- Providing scientific and technical advice and assistance to the Military Committee in the field of aerospace research and development (with particular regard to its military application);
- Continuously stimulating advances in the aerospace sciences relevant to strengthening the common defence posture;
- Improving the co-operation among member nations in aerospace research and development;
- Exchange of scientific and technical information;
- Providing assistance to member nations for the purpose of increasing their scientific and technical potential;
- Rendering scientific and technical assistance, as requested, to other NATO bodies and to member nations in connection with research and development problems in the aerospace field.

The highest authority within AGARD is the National Delegates Board consisting of officially appointed senior representatives from each member nation. The mission of AGARD is carried out through the Panels which are composed of experts appointed by the National Delegates, the Consultant and Exchange Programme and the Aerospace Applications Studies Programme. The results of AGARD work are reported to the member nations and the NATO Authorities through the AGARD series of publications of which this is one.

Participation in AGARD activities is by invitation only and is normally limited to citizens of the NATO nations.

Published February 1991

ISBN 92-835-0604-9

Printed by Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ

Theme

In the early 90s the ubiquitous availability of broader communication bands with ISDN, for example (Integrated Service Data Networks), and other channels, such as electronic mail, the availability of easy access and user friendly protocols, the processing of natural languages with the help of artificial intelligence, the use of machine-aided translation (computer assisted or automatic translation) will have many effects on information retrieval and means of communication.

These new concepts will result in better fulfilment of the needs of users, who consider the delay in information delivery as a more and more important factor.

Our information systems and centres must be prepared to make full use of these new technologies, which should enable a better fulfilment of user's needs. The final paper proposed short-term actions to be undertaken and recommendations to be transmitted to authorities and all persons responsible for information systems and centres.

Thème

Au début des années 1990, l'omniprésente disponibilité de bandes de télécommunications plus larges, avec, par exemple, les RNIS (réseaux numériques à intégration de services), parmi d'autres voies de transmission telles que le courrier électronique, la disponibilité de protocoles conviviaux et d'accès facile, le traitement du langage naturel à l'aide de l'intelligence artificielle, et la mise en œuvre de la traduction machine (traduction assistée par ordinateur ou traduction automatique) auront des répercussions importantes sur la recherche documentaire et les voies de communication.

Ces nouveaux concepts permettront une meilleure satisfaction des besoins des utilisateurs, lesquels attachent de plus en plus d'importance à la réduction des délais dans l'acheminement de l'information.

Nos systèmes et nos centres d'information doivent être en mesure d'exploiter au maximum ces nouvelles technologies, et de ce fait, assurer une meilleure satisfaction des besoins exprimés. La dernière intervention proposé des ıctions à entreprendre à court terme, ainsi que des recommandations à transmettre aux autorités compétentes et à tous ceux qui sont responsables de systèmes ou de centres d'information.

Technical Information Panel

Chairman: Mr Albert Yanez
Conseiller du Directeur
C.E.D.O.C.A.R.
00460 Armees
France

Deputy Chairman: Mr Richard Searle
Chief Librarian
Royal Aerospace Establishment
Farnborough
Hants GU14 6TD
United Kingdom

MEETING PLANNING COMMITTEE

Mrs Randi A.Gjersvik (NO) — Director
Mr Oleg Lavroff (FR) — Vice Director

Lt. Col. A.Cuffez (BE)
Mr B.Hisinger (DK)
Mme F.Lhullier (FR)
Mrs D.Patrinou (GR)
Mr G.Breas (NE)
Mrs F Urundul (TU)
Mr R.Searle (UK)
Ms B.Lawrence (US)
Mr K.Molholm (US)
Ms C.Walker (STC)

PANEL EXECUTIVE

Mr G.W.Hart

Mail from Europe:
AGARD—OTAN
Attn: TIP Executive
7 rue Ancelle
92200 Neuilly sur Seine
France

Mail from US and Canada:
AGARD—NATO
Attn: TIP Executive
APO New York 09777

Telephone: 33 (1) 47 38 57 95
Telex: 610176 (France)
Telefax: 33 (1) 47 38 57 99

Contents

	Page
Theme/Thème	iii
Technical Information Panel	iv
	Reference
Keynote Address by H.K.Johansen	K
SESSION I – DEVELOPMENTS IN THE PREPARATION, TRANSFER AND COMMUNICATION OF DOCUMENTS (TECHNIQUES USED, NETWORKS AND COMPATIBILITY OF SYSTEMS) **Chairman: Mrs R.Gjersvik (NO)**	
An Overview of Optical Media, Now and in the Future by R.Storleer	1
Computer-Assisted Publishing with its Standards and Compatibility Problems by H.Wendt and C.Zürn	2
Broader Communication Bands by G.Stette	3
SESSION II – PROCESSING OF NATURAL LANGUAGES: APPLICATIONS OF ARTIFICIAL INTELLIGENCE (EXPRESSION OF A NEED, INFORMATION RETRIEVAL, COMPUTER-AIDED TRANSLATION) **Chairman: Mr A.Yanez (FR)**	
Fields of Application of the Processing of Natural Languages with the Help of Artificial Intelligence by B.Maegaard	4
Full Text Retrieval with Graphics by M.Lesk	5
Traitement du Langage Naturel: Présentation d'Applications Pratiques par N.Benamou et L.Canivez	6
Beyond Machine Translation by L.Rolling	7
The Role of Intelligent Online Interfaces to Bridge the Communication Gap by A.Vickery	8
Paper 9 withdrawn	
SESSION III – DATABASES, DATABANKS OF THE FUTURE (PROCESSING OF FULL TEXT) **Chairman: Ms B.Lawrence (US)**	
Database Publishers' Challenges for the Future by B.Lawrence	10
Exemples de Bases de Données Utilisant des Textes Numérisés (Textes Intégraux ou Résumés) par P.Pellegrini et P.Laval	11

Reference

The Economic Aspects of Developing and Marketing Full Text Databases 12
by M.Hepworth

SESSION IV – IMPACT ON INFORMATION SYSTEMS OF THE AEROSPACE AND THE DEFENCE FIELD, OF DOCUMENTATION CENTRES AND ON THEIR USERS
Chairman: Mr O.Lavroff (FR)

Bridging the Communication Gap: The Case of the Portuguese Information System for Industry 13
by M.J.Barrulas and Z.Correia

Possible Recommendations to Ministries and other Authorities based on the Foregoing Papers 14
(Projets de Recommandations à Faire auprès des Ministères et auprès d'Autres Autorités sur la Base des Communications Présentées lors de la Conférence)
by O.Lavroff

Attendance List A

KEYNOTE ADDRESS

by

Henry Kjell Johansen
Chief Scientist
Norwegian Defence Research Establishment
P.O. Box 25
N-2007 Kjeller
Norway

Mr Chairman, ladies and gentlemen,

It is a great pleasure to welcome all of you to Trondheim and to this AGARD TIP specialists' meeting on "Bridging the Communication Gap".

I am very happy to see that so many of you have found the way to Trondheim, the centre of Mid-Norway. Geographically, it is not in the middle of the country, only about ⅓ of the way up to North Cape, the most northern part of mainland Norway. If we also include Spitzbergen island, where we have a couple of Norwegian towns, it is only ⅕ of the distance up to the most northern point. Still, Trondheim is at the same latitude as Fairbanks in Alaska or Godhope in Greenland.

To illustrate the length of our country, if we could turn Norway around its southern point, North Cape would reach down to Rome and the frosty Spitzbergen would be placed in the middle of the Sahara desert in Africa. In this long and impractical country we have only 4 million people.

Historically and culturally Trondheim has been the centre of Norway. It was founded by the Viking King Olav Tryggvason in the year 994 ie 1000 years ago. You will see a monument to the King in the middle of the city. Trondheim was also the capital of Norway during the Viking Age. It is now the third largest city in Norway.

It is not because of its past that we have organised the TIP meeting in Trondheim. It is the leading city when it comes to higher technical education and research, with its Norwegian Institute of Technology and the SINTEF research organisation. The Technical University Library of Norway here in Trondheim is the leading library in the country in respect of the development and exploitation of modern computer based library services. In the early 70s the library started developing their BIBSYS system, and the latest version of that system has now become a standard for research libraries in Norway.

AGARD's main task is to stimulate the exchange of scientific and technical information in the aerospace field in order to help strengthen the NATO defence posture. Last year the AGARD technical panels, with their 450 scientific and technical experts, sponsored more than 40 conferences, which attracted a total attendance of about 6000. The panels of AGARD organised 60 working groups and produced 80 publications. 90 support projects to Greece, Portugal and Turkey were established.

How do AGARD's admirable work and goals comply with the revolutionary political changes in Europe where Eastern Europe is liberating itself and the Soviet Union has embarked on the long journey toward a free society? The London Declaration on a Transformed North Atlantic Alliance issued by the North Atlantic Council in July this year opens up NATO as an alliance of change. It will continue to provide for the common defence, but NATO should also be an institution where Europeans, Canadians and Americans work together not only for the common defence, but to build a new partnership with all the nations of Europe. The Alliance's integrated force structure will change fundamentally to include the following three elements:

- NATO will field smaller and restructured active forces. These forces will be highly mobile and versatile to allow for maximum flexibility in responding to a crisis.

- NATO will scale back the readiness of its active units, reducing training requirements and the number of exercises.

- NATO will rely more heavily on its ability to build up larger forces if and when they might be needed.

To reduce the Alliance's military requirements it is also essential to have sound arms control agreements with effective arms reduction verification regimes.

Where do these major transformations bring us with respect to AGARD? The AGARD NDB has set up a group of Senior National Delegates to work out possible scenarios and propose changes. We do not have the report from that group yet, but some of the more important considerations would be the following.

- The expenditures on military defence in the NATO countries will decrease. Some countries have already announced deep cuts of the order of 20% over the next 5 years.

- The restructuring of our forces, and higher mobility, can not be achieved through organisational measures alone. We have to develop new cost-effective equipment suitable for the new tasks.

- A reduction of training requirements to allow lower readiness and fewer exercises will also require new equipment which is easier to operate and which can be stored for years without being used.

- The ability to build up larger forces requires the capability to allow a sudden increase in military production.

All these partly conflicting requirements will require greater effectiveness of military R&D and production. To reach these goals the output from the NATO R&D community has to increase without a corresponding increase in expenditure. The way ahead includes measures such as:

- Putting more effort into the reduction and avoidance of unnecessary duplication of R&D work within the NATO countries.

- Better coordination between military and civilian R&D. In many areas this means that military R&D should be added onto civilian R&D programmes.

- More standardisation of equipment and support systems.

A more efficient way of managing the scientific and technical information flow within the countries as well as between the countries is a key issue in order to achieve these goals and may be one of the most challenging tasks for NATO. When we know that over 6000 scientific articles are written each day and a doubling is expected over 6 years we all understand how demanding this task is, namely to get the right information to the right person at the right time. Improving the management and dissemination of technical information could reduce the duplication of R&D effort in the Alliance, improve the quality of work done and thereby reduce the total expenditure on defence.

AGARD does not have a direct responsibility for standardisation in NATO, but to achieve the NATO goals, standardisation is mandatory when it comes to reducing the cost of equipment and support systems. Indirectly, AGARD plays an important role in standardisation through the work in specialists' meetings and in the panels. Common understanding of the technological possibilities, common terminology and exchange of information about ongoing research programmes have in the past led to common approaches to solve problems and paved the way for standards.

Seen in the perspective outlined, this TIP specialists' meeting on "Bridging the Communication Gap" is highly relevant and timely. We now see the technological possibilities to realize effective methods for handling the information flow. The greatest difficulty will be to get acceptance for common standards which can ensure simple and effective exchange of information within and between the NATO countries, and between the civilian and the military communities.

AGARD meetings make it easier for people from the different organisations and countries to get together and exchange views and information at the conference and in a more informal manner outside the conference room. To establish new friendships and contacts is one of the purposes of arranging AGARD conferences.

The name Trondheim comes from the old Norwegian "Trondheimr", which means "The home where you thrive". I hope that you all will feel comfortable here and that you enjoy your stay in Trondheim. I wish you a successful meeting.

AN OVERVIEW OF OPTICAL MEDIA NOW AND IN THE FUTURE

by
R.STORLØER
Senior Research Librarian
THE TECHNICAL UNIVERSITY LIBRARY OF NORWAY
DOCUMENTATION DEPARTMENT
Høgskoleringen 1
N 7034 TRONDHEIM NORWAY

SUMMARY

This paper deals with the Optical Technology Family where the different members of the analog, hybrid and digital optical storage groups are presented. So much happens so fast in this area and the users are confused concerning which media are useful for what.
The different media have both advantages and disadvantages.
Analog storage has up to the present been the only medium for storage and presentation of pictures and films with high quality (eg. LaserVision, LV-ROM).

By digitalization of pictures, each picture needs from 0.3 - 1.0 Mbyte depending on the quality. That means that a CD-ROM only can keep about 7-900 color pictures and only 25-35 seconds film presentation.
A videodisk can keep more than 100 000 pictures which gives about one hour film presentation.

The involved producers are working hard to solve this problem by finding new comprimation algorithms.
The presentation will deal with the different types : LaserVision, RID, WORM, DRAW, CD-ROM, CD-ROM-XA, DVI, CD-I, Erasables etc.

What about storage capacity, which type of matter is useful to store on what media, existing standards, use and miscellaneous, future prospects etc. Main emphasis is drawn to CD-ROM which is the most common used optical media in a library environment.

INTRODUCTION

A PC can handle electronicly stored data. This has given publishers and information intermediaries possibilities to distribute their products by way of electronics.

Magnetic storage has since the computer technology entered our working environment been the dominating method in storing data, but this medium has in fact the later years got competition from new storing technologies as optical storage devices due to these medias excellent durability and enormous storing capacity.

Optical storage makes it possible to store text, pictures, sound and data at relative low costs, and with fast access to the information.
The medium is small and light, and gives the user access to enormous quantities of information.

HISTORY

By the late 1920s, encoded optical storage were utilised by the recording of motion picture soundtracks. In the recording process, the sound is converted into an electrical signal which modulates a light source that exposes the sountrack portion of the film. This produces a squiggly line on the film. This method of storing sound information is analog optical storage and is still used for motion picture soundtracks.

In the 1940s and 1950s, the process of magnetic recording was developed and came into widespread use. This technique involves storing information as a magnetic pattern on a moving magnetic surface; a time-varying electric current passes through a magnetic core, producing a corresponding time-varying magnetic field.

The digital information used by computers is stored as patterns of 1s and 0s, or ONs and OFFs. The encoded data does not suffer from the degradation that can affect analog information. Unlike an analog recording, digitally stored information is made up of only two signal levels: ON or OFF. A copy is not merely an approximation of the digital source, but an exact bit for bit reproduction.

ANALOG AND DIGITAL OPTICAL STORAGE

14 years ago the dutch company Philips demonstrated its first videodisk, VLP, Video Long Play and in 1981 Philips put the disk on the market under the more wellknown name LaserVision. Since then the development in this area has moved very fast foreward.

At the moment there are about 100 different members of the optical storage family. Most of them are unknown and have appeared on the market the last 2 - 3 years, and some of them are still in the lab.
With so many different media to choose between, it can be difficult to choose the right type for a special purpose.
Only a few of all the different optical storing devices today are in real commercial use.

The optical family can be divided into two main groups, the reflective and the transmissive group. This presentation will deal with the reflective group..

THE TECHNOLOGY

There is some differences between the different disks, but they have some common characteristics. The disk which is mainly

produced from plastic is 1 - 2 mm thick and consists of different layers. The innermost layer has a concentric track of different length depending on disktype. This track consists of microscopical pits with different size and distance.

On analog disks the size and distance between the pits depend on the information to be stored. The reality is that the information stored on an analog disk is stored in a digital manner, but the reflected laser beam acting as the information carrier is after passing through the photo decoder, decoded to an analog electrical signal.
Typical for the analog signal is that it is a time-varying signal.

Digital storage of information on the archival disks (>20 cm diameter) may be represented as holes and "nonholes". Holes are 0s or OFFs, and "nonholes" is or ONs. The size of the holes are exact and the distance between the holes and "nonholes" are the same all over the disk. The reflected laser beam represents a continous stream of ONs and OFFs which after the photo decoder is representing an electrical signal which then is decoded into ONs and OFFs by a high technical and complicated process. Since the computer works with digital signals (ONs and OFFs), we do not need any mechanism for converting the signals from digital to analog.

To have an as effective as possible storing of digital information, the digital compact disks were developed. They are single sided disks and the binary numbers 1s and 0s are not represented by the low reflecting holes and the high reflecting "nonholes".
It is the change between pit/land or land/pit that gives the binary value 1. The pits and the distance between the pits vary and give the respective 0s (OFFs).

The information stored on the compact disks are therefor three times more close together than on the big archival disks.

The next layer is an extremely thin layer of polished aluminium which follows the pits of the innermost layer and acts as the reflective part of the disk. The outermost layer is a protective plastic coating of polycarbonate. Fig. 1 shows the cross-section of a compact disk.

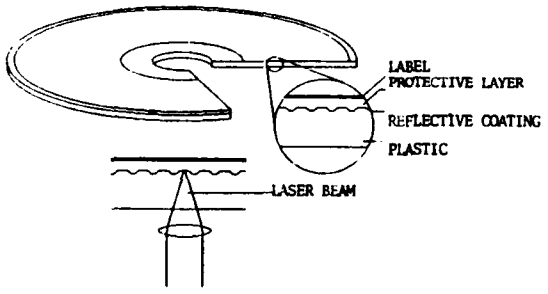


Fig. 1 Compact Disk Cross-Section

An intensive and accurate laser beam in the optical reader hits the aluminium foil of the disk and is reflected back to a photo cell where the signals are converted to electrical signals.
Because of the different size and distance of the pits, the reflected laser beam will differ in intensity and wavelength from time to time and the photocell will recieve different signals (Fig. 1).

Constant Angular Velocity (CAV) and Constant Linear Velocity (CLV)

CLV

Before we start presenting the different main members of the optical family, we have to explain two terms. The information can be stored at the same density all over the disk. That means: the disk reader has to change the rotation speed depending where the laserhead is reading information on the disk, and the disk has to rotate with a constant linear velocity (CLV).

CAV

When the information is stored with higher density near the centre than in the periphery of the disk, the disk will rotate at the same speed independent of where the laserhead is reading information. We say that the disk is rotating with a constant angular velocity (CAV).

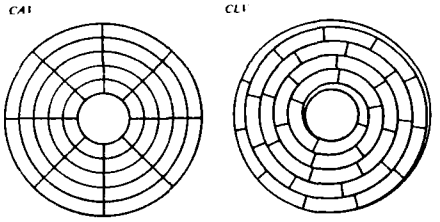


Fig. 2 Sector structure on CAV and CLV Disks

THE REFLECTIVE OPTICAL FAMILY

The reflective optical family consists of tree main groups; the analog, the hybrid and the digital groups and each group contains many members (Fig. 3).

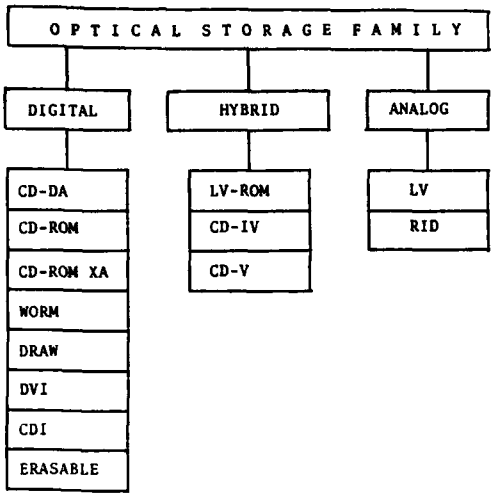
THREE STANDARDS

PAL (Phase Alternation Line) is the european TV standard giving 25 pictures pr. second.

NTSC (National Television Systems) is the american standard giving 30 pictures pr. second.

SECAM (Sequentiel à mèmoire) is the french standard and is also used in East-Europe.

A PAL disk can not be used on a NTSC reader and vice versa but today socalled dualplayers or combiplayers are on the market which can handle more than just one standards.



CD-DA	: Compact Disc - Digital Video
CD-ROM	: Compact Disc - Read Only Memory
CD-ROM	: CD-ROM Extended Architecture
WORM	: Write Once Read Many
DRAW	: Direct Read After Write
DVI	: Digital Video Interactive
CDI	: Compact Disc Interactive
ERASABLE	: Erasable Disc
LV-ROM	: LaserVision - Read Only Memory
CD-IV	: Compact Disc - Interactive Video
CD-V	: Compact Disc - Video
LV	: LaserVision
RID	: Recordable Interactive Disc

Fig. 3 Some of the members of the Optical Storage Family

THE ANALOG DISKS

LaserVision

The videodisk or the more wellknown name LaserVision was introduced on the market in 1981 by Philips and succeded the VLP - Video Long Playing disk.
This disk is quite suitable for storing pictures, movies (video) and sound at high quality. The disk with a diameter of 30 cm contains 54 000 tracks on each side for storing pictures and two soundtracks for stereo sound or possibilities for bilingual sound.

Each track can be accessed randomly, which makes it possible to acess an exact picture in very short time. The entire disk can be scanned in 25 seconds.
This is an advantage compared with the videotapes where we need much more time to find an exact picture sequence.
The videodisk is an advanced extension (version) of the videotape and was original developed for the entertainment industry.
The disks exists as bot CAV and CLV disks.

LaserVision-CAV

On the CAV disk one picture is stored on each track, each one with an unique address. The disk rotates with a constant angular velocity. A disk in PAL format can keep 36 minutes video or 54 000 single pictures with stereo sound on each side. Because of the unique addressing of each picture, the disk is well fitted for presenting single pictures (dias) or short film sequences.

The disk can be contolled from a remote controller or a PC/computer and is therefor an interactive system, but the data on the disk can not be treated by a computer without special applications.
The disk is produced for all three standards PAL, NTSC and SECAM.
The primary market for the disk is schools, education, training, advertising etc. Under normal conditions, the disk rotates at 1 500 rpm.

LaserVision-CLV

The disk exists for all three standards (PAL, NTSC and SECAM) and is quite similar the CAV disk apart from that the disk rotates at constant linear velocity and therefor can hold 2x60 min video with stereo sound.
Under normal condition in PAL format the rotational speed varies from 1 500 - 570 pr minute.
The disk is suitable for storing movies and is aimed at the entertainment and consumer market, but the disk appeared a little to late on the market, since the videocassette already was established with a very large market share.

LV-CAV-P and LV-CAA

Two other products in the LaserVision family is the LV-CAV-P disk where P is the acronym for Professional and the LV-CAA disk where CAA is the acronym for Constant Angular Acceleration.

The first one may contain 36 min video or 54 000 single pictures, stereo sound and 350 Mb information on each side and the second 2x2 700 single pictures and 2x22,5 min video with stereo sound.
The disks are used for automatic information and sale applications (point of sales and point of information).

The disks are also used for presentations on multivision or videowalls with 16, 24 or 30 TVs placed together. All the screens can present the same picture, parts of a greater picture or pictures with different configuration.
Neccesary equipment is a computer to distribute the different pictures to the different screens.
LV-CAV-P appeared on the market in 1981 and LV-CAA in 1987.

RID-CLV and RID-CAV

RID is the acronym for Recordable Interactive Disk and both the CAV and CLV disks appeared on the market in 1987. The CLV disk can keep 2x60 minutes video with stereo sound and the CAV disk 2x54 000 single pictures or 2x30 minutes video with stereo sound. Both disks are for local recording of video and the CLV disk also for local recording of still pictures. They are useabel for local recording in connection with marketing, sale and education.

Recordable videodisks open up for big possibilities since the quality is much better than for videotapes.
The CAV disk is also suitable for multivision or videowalls.

SOME COMMENTS

The videodisks were developed for the storing and playing of video material as movies and audiovisual information for education, entertainment, marketing and sale.
The disks can also contain digital data. The digital information is then encoded on a videosignal and stored on the disk in the same manner as an analog signal, but there is some difficulties since the pits are very small (1 micrometer) and sensitive for dust particles and irregularities on the disk.
Since the pictures on a videodisk is stored as analog signals they can not be treated or manipulated by a computer without special equipment. To day we can install an interfacecard in our PC (eg. MIC 4000), which makes it possible to convert analog stored pictures on the videodisk into digital data for PC use. But the storage requirements for an analog picture in good TV quality is from 1/2 - 1 Mb in digitalized form. That means for a 600 Mb compactdisk only 30-35 seconds of video (PAL).

THE HYBRIDS

This medium is based on analog storing of pictures or video while the sound or text are digitalized. Therefor the name hybrid, it is a bastard since it has the possibilities for both analog and digital storing.

CD-V - Compact Disk - Video

The disk uses analog video and digital sound. CD-V is based on the standard described in **THE BLUE BOOK** (Philips) and was introduced to the market in 1988. The standard defines three different disk sizes, 12 cm (CD-V-S Compact Disk Video Single), 20 cm (CD-V-EP Compact Disk Video Extended Play) and 30 cm (CD-V-LP Compact Disk Video Long Play).

All different sizes can be played on socalled "comiplayers".
The use of analog video means that we have to take into consideration the three standards PAL, NTSC and SECAM.

The 12 cm CD-V has a golden surface to distinguish it from the more famous relative CD-DA (Compact Disk - Digital Audio) of the same size.
Each disk can contain 9 000 single pictures or 6 min video (PAL) and 20 min stereo sound (HiFi).

The EP and LP disks are dobblesided and keep 20 and 30 min stereo sound (HiFi) on each side.
The disks are directed to the consumer market.
Combiplayers can handle CD-DA (digital music), CD-V-S, CD-V-EP, CD-V-LP (digital music + TV) and LV-CLV (analog music + TV).

LV-ROM - LaserVision Read Only Memory

The disk is for interactiv use and was introduced on the european market in late 1986 and keeps about 100 000 pictures (PAL), sound and data. The pictures are stored in analog form while the sound and text are digitalized.The disk is useful for educational and instructional purposes, and the users are both from schools, universities and private companies.

One of the more known projects where LV-ROM is used is **THE DOMESDAY** project. The Domesday Project commemorates the 900th anniversary of the Domesday Book, compiled on the orders of King William I. However instead of parchment, the 1986 version uses LV-ROM technology, which incorporates digital data on a LaserVision disk. The database for the project is contained on two disks, a national and a community one.

CD-IV - Compact Disk Interactive Video

A 12 cm disk contains 600 Mb of data. It is a combination of CD-I and CD-V technology and can store entertainment, facts and educational information.
CD-IV is a detached system with it's own CPU which is hooked up to a TV, and controlled with a remote control.
Time for presentation is planned to 1990/91.

THE DIGITAL SYSTEM

History

Philips was the pioneer within the area of digital optical storage. They started experiments in the early 70s. These early experiments lead to the Megadoc system which was introduced in 1978 and uses a 30 cm optical disk that contains 3 Mb of data. Information such as text, pictures or sound are converted to digital form, and stored as binary codes (pits and lands) on the high reflecting optical disk.

CD-DA Compact Disk Digital Audio

The Red Book

Although the laserdisk did not prove to be a success in the consumer market, it led directly to the development of the digital audio compact disk, the largest success to date within consumer electronics.

Philips and Sony are competitors and hard working in the optical storage media area. But they do cooperate through an agreement to share their optical disk technology in order to develop a standard for optical recording of digital encoded music specified in The Red Book of june 1980.

In late 1982 they began to market audio compact players. The standard describes how the data (music) should be organized on the disk (CD-DA). It is the same concept as applied for the CD-ROM except that much higher demands are made regarded on miscorrection and coding of the data.

The CD-DA players were an almost immediate success and the consumers eagerly accepted the medium because of its overwhelming advantages.
Some figures show this explosion by 35 000

players sold in 1983, 200 000 in 1984, 800 000 in 1985 and about 50 millions in use in 1990.

The 12 cm CD-DA disk can keep 72 min HiFi sound and 32 Mb data, text and pictures and the disks are singlesided.
The CD Audio disk carries an exact, digital copy of the master recording and has an extremely high sound quality.

WORM - Write Once Read Many (Times)

On this medium we can store the information by ourselves, but already stored information can not be changed or deleted. Stored information can be reused as many times you want.
The most commercial WORM disks have one or more vacuum-stored metal layers where the information is stored. The storage of information is done by a laser beam in the metal layer either by holes, bubbles or by melting some of the layers locally together.

The disk is very stabil even after long time of storage, and the durability does not decrease even after frequent use. The first generation WORM disks were byproducts from the videodisks. The 12" disks became commersial available in 1983 and could keep 1 000 Mb (= 1 Gb) of information on each side.

An example on the WORM disk is the Megadoc system from Philips. On the Megadoc disk we can store 1 Gb of information on each side. This can be illustrated by 500 000 typewrited pages of text - equal to one milliard keystrokes or about 170 shelfmeters.

The second generation WORM readers became available in 1985. They are smaller and cheaper than the first generation readers and use 5 1/4" disk size. The readers are available as external and internal models.

All type of information can be stored on the disk. Examples are archiving of documents (patientjournals, contracts, letters, bills, insurence agreements, banking information etc), drawings, photos, tables, statistics etc. Each document gets it's own unique address on the disk and can easily be retrieved.

DRAW - Direct Read After Write

This disk which is also named E-DRAW (Erasable Direct Read And Write), is quite similar to the WORM disk and is used for storing the same type of information, but differs in some ways. If we recognize somthing wrong with information already stored, these parts of the disk can be marked as unuseable and the correct information stored. The laser will ignore the market areas on the disk. The disks which were introduced in 1984, exits as both 5 1/4" and 12" with storing capacities from 300 to 1 000 Mb on each side.

CD-ROM - Compact Disk Read Only Memory

The High Sierra Group Format or The Yellow Book

As an exstension to the CD-DA specification mentioned in The Red Book, Philips and Sony announced, a new specification for data storage in october 1983.

In november 1985 companies working in the optical media area came together near Lake Tahoe in High Sierra to make agreements for a standard for the new optical media CD-ROM.
ROM is the acronym for Read Only Memory and means that the disk may only be read and not written on by the user.
Representatives from Apple Computer, Hitachi, Philips, Microsoft etc. named themselves the High Sierra Group, after their first meetingplace.
Their common agreement became the specifications mentioned in The Yellow Book. The defacto standard became later ISO 9660.

The CD-ROM

A CD-ROM is a 12-cm-diameter CLV disk which can keep more than 550 Mb of data corresponding to 275 000 typewritten A4 pages or 1 600 floppy disks. The information is stored in a five km concentric track on the disk.
The CD-ROM has a digital output signal and does not need any digital/analog convert mechanism.

CD-ROM is originally developed for storing textual and graphical information, not sound like CD-DA and therefore it needs a much better error correction system for the encoded data than CD-DA.
Beside CD-DA, CD-ROM is the only digital optical storage media that has become a hit and it is specially within the library world the users are found.

The CD-ROM is aimed to the library, school, agencies and business market.
Since the first CD-ROM databases were bibliographical or reference databases having a corresponding online database or a printed abstract journal, it was natural that the libraries were the first and are still the largest users of the new medium.

Compared with magnetic disk storage the advantages with CD-ROM are quite obvious; Because it is no contact between the laserhead and the disk, the CD-ROM disks do not suffer from head crashes or physical wear.
The medium itself is extremely durable and difficult to damage and there is as yet nothing to suggest that they have a limited life-span.

CD-ROM is appropriate for the distribution of large amounts of data. This could be electronic publishing of databases, maps, library catalogs, abstract journals, books, magazines and periodicals, market and finaciel information, parts list, repair manuals, historical data and other material that is typically distributed through online services or on paper.

CD-I - Compact Disk Interactive

The Green Book

CD-I is like CD-DA and CD-ROM a result of the cooperation between Philips and Sony and in february 1986 they announced the standard for CD-I.
The specification refered to as The Green Book describes methods for providing audio, video, graphics, text.

The specification for CD-I is as strict as for the CD-DA concept and is intended to provide a world standard for complete crossbrand compatibility of players and software.
When I visited Philips in august 1989 they were ready to introduce the CD-I equipment to the market.

The 12 cm CD-I disk can keep 650 Mb of information. These can be used for 16 channels speech sound in mono, 8 channels speech in stereo, 8 cannals music, 4 channels music in stereo, 4 channels HiFi music, 2 channels HiFi music in stereo or 1 channel CD-DA audio in stereo. Each channel can give 70 minutes playing time; that gives a total of about 19 hours of mono speech.

Depending on the quality, the number of pictures stored on one disk differ from 1 000 pictures in high quality and complete color possibilities (256 colors) to 20 000 pictures with more simple graphic and few colors.

One of the main disadvantages with CD-I is that it is using it's own CPU and is a single station which cannot be connected to a PC, but like a CD-DA player the CD-I unit can be connected to a TV set, videoplayer or stereo equipment.

Potential CD-I applications are entertainment, education, training, in-car navigation, home shopping and reference tools.

The CD-I concept has not yet solved the problems for full screen, full motion video (motion pictures) presentation, but full-screen frame-based animation is possible.

CD-ROM XA - Compact Disk Read Only Memory Extended Architecture

CD-ROM XA is a 12 cm disk equal CD-ROM and a result from the cooperation between Philips, Sony and Microsoft. The first release of CD-ROM XA was announced in march 1989. The CD-ROM XA specification makes it possible for an application to play audio information at the same time that it retrieves other interleaved information such as text, images and graphics from the CD-ROM disk.

With CD-ROM XA, audio signals are compressed to take up less space on a disk. By using different levels of compression, one disk may contain up to 16 hours of mono quality sound. But even when using stereo FM quality audio, one disk can still contain 4 hours of sound.

The compression technology which is defined in the existing CD-I specification (The Green Book) is called ADPCM (Adaptive Differential Pulse Code Modulation) and will allow disks to be coded with compressed digital audio information in an interleaved fashion.
Interleaving is an important part of the standard, and allows for up to 16 separate data tracks to exist in each sector. In this way the application can switch languages (audio and/or text) instantly by simply reading different relative blocks in the same interleaved file.

Time-span for detailed specification is as follows:

Step 1 : Text and sound (CD-ROM + ADPCM) march 1989

Step 2 : Single presentation of images (VGA,CD-I) end 1989

Step 3 : Full-screen presentation of film 1991

The new CD-ROM XA format will allow publishers to create disks that will be playable not only on any suitably equipped PC but also on any CD-I system.

Most of the criticism around CD-I is related to the fact that it was based on a specific operating system and hardware environment and hence would be frozen in time and inflexible in the future. CD-ROM XA would appear to go around this problem for those publishers and users who want the capabilities of CD-I but in a general PC environment rather than in a consumer hardware specific environment.

CD-ROM XA is a bridge between CD-ROM and CD-I and is aimed at mainly professional and institutional markets and reference type applications. The CD-ROM XA system is basically transparent for any PC and independent of a particular CPU and operating system.
CD-ROM XA is just an ordinary CD-I for the PC-world.

DVI - Digital Video Interactive

The DVI concept was developed at David Sarnoff Laboratories of RCA in Princeton, New Jersey. DVI is a cooperation between IBM, Microsoft and Intel. The prototype has existed since march 1987 and integrates graphics, audio and video images with text.

The DVI protocol involves processing about 22 Mb or 12.5 millions calculations per second of text data, sound and digitally encoded video images. At this rate of processing, a 550 Mb optical disk would hold less than a minute of data.
That is why the the protocol involves introduction of new tecnology to compress data by a factor of 150 to 1 and to decompress it so fast that a 30-frame-per second (NTSC) image can be transfered from the disk to a video monitor in real time. This makes it possible to get about an hour of video on a small disk.

In a typical application one of these disks might combine 20 minutes of video with 15 000 pages of text, 5 000 high resolution

still images and six hours of audio.

The DVI disk is 12 cm diameter and the equipment is to be connected to a PC. The DVI concept is build up around two microchips placed in the PC. As mentioned above the pictures are comperessed by a factor of 150. For the compress process two video display processors (VDP) chips with VLSI (very large scale interface) architecture are used. One chip handles the compress algorithm and permits not only singel images but motion/video film, text and graphic in the same image. The other chip handles the dissolution of the image, screen size and the interface against the PC. The concept is based on storing the canges in the image.

The producers hope that DVI will replace films, video casettes, videodisks and different compact disks.

ERASABLE DISKS

The 12 cm erasable or WARM (Write And Read Many (Times)) disks to day are based on magnetooptical storage of the information where a thin magnetic layer covers the reflective layer of the disk. The needles in the magnetic layer is placed horizontal in the plane of the disk.

An intens laser beam heats a part of the disk while a strong electric coil change the magnetism in that part.
The heating continous until the Curie-point or Curie temperature is reached. The Curie point is the temperature where a ferromagnetic material by heating looses its ferromagnetism and becomes paramagnetic. This happends at 760'C for iron.

The whole concept is based on the KERR effect saying that changes in the polarization condition for light (laser) which is reflected from a ferromagnetic material (iron, cobolt,nickel) depends on the magnetic condition of the matter.

To write information on the disk, a special laser reverses the magnetic direction of each predetermined spot on the disk surface. As the polarity is switched, the differences are read as 0's and 1's bits of information. A low-intens laser is used to read the information on the disk without changing the polarity of the spots.
The disks can keep from 300 to 650 Mb data. Erables, just like WORM systems, lack standards in the interfaces, size of disks and drive system, and in the lack of drivers, or software for interfacing the optical drives with microcomputer workstations.

Another problem for WORM and erasables has been the slower access time and data transfer rates when compared to Winchester hard drive systems.
To day's erasables have 60-70 ms access time and 5 Mbits per second data transfer. With increased production and sale, the cost for drives and disks will decrease.

OTHER SYSTEMS

In one system the information can be stored on a small "credit"-card which is read by a laser.
In another system the information is stored on a paper band which is read by a light beam. Paperstored information needs much more room than the other systems. The bar codes used on commodities in the stores contain 13 chiffer values describing where the commodity is produced, who is the producer and name and price of the commodity.

In some systems the information is stored on tape casettes. A laser reads the information on the tape. One of the disadvantages with the tape medium is that it takes more time to retrieve the information than on a disk.

THE FUTURE

This presentation has dealt with just a few but mainly the most important members of the optical media family, and I have focused on the members of the reflective part. The number of members in this family have increased rapidly the last few years. In most of the systems the information is stored on reflective disks as pits and lands, but other optical storing methods exists.

Many of the products have been introduced to the market with entusiasm and go-ahead spirit, but since then nothing is heard about them. They are here today and gone tomorrow. Some of the products have never reached the market, but have ended it's days in the lab.

The big industries in the optical field have a lot of development cost and hope to make the big score.
The problems the industry has to face is that the technology is suffering from lack of standards.

Analog storage for film presentation will exist untill the digital technology can handle the same matter. To days compression algoritms are not good enough compared with analog storage.

CD-DA is the only product that has became a hit with almost 50 million produced players. The reason for that success is the exact and well defined specification from the beginning which became a worldwide standard. In the video tape world we still have the fight between the Beta vs. VHS standard and in the videodisk world we have the different noncompatible products like LaserVision, SelectaVision and VHD.
Every audio compact disk is compatible with every CD player in the world.

On the digital side for multimedia presentation the fight will be between CD-ROM XA, CD-I, DVI and the erasables and it is difficult to predict a winner.
All of the products contain technologies and will get further technologies that will make them of big interest in the future.

New compression algorithms will solve the full screen film presentation problem, but

I think the hardware and software producers will sit on the fence until standards have been established. These industries have major possibilities, but cannot really take off until standards have been established. Many of the products are of great interest, but will not go commercial because no one wants to risk large amounts of money in a technology that may not succeed.

The industry is also suffering from a related problem of vision. The CD-ROM and multimedia industry must know where it is going and on which markets these new technologies would be used, what form they would take, or even what they would be used for.

Untill the uses, formats and economies becomes established, it will be difficult for most libraries to just purchase media that each require incompatible extensive hardware platforms. Libraries are often the environment where multimedia are used and act as trend setters for multimedia workstation.

The CD-ROM multimedia industry will be a major area of growth and competition in the nineties. Several "standard wars" which is no unfamilar phenomena in the computer world have begun and as always in such fights it is realy difficult to predict the result.

According to a report by Frost & Sullivan Inc. called the Optical Memory Market in Western Europe the optical storage market is expected to soar by 1993. The market for disk drives and disk media will increase from $92 million in 1988 to $913 million a year by 1993, West Germany buing the most, followed by Britain and France.
The small erasable disk drives are predicted to become the dominant form of optical storage, accounting for 72% of 1993 optical disk drive dollars.
In media, the small write-once disks will take 85% of the market.

RECOMMENDATION

For those of you that want to read more about the historical and technological aspects of digital optical media, i will recommend The Brady Guide to CD-ROM.
For those of you who are interested in capacity figures and use for both analogs and digitals, and handle the sweedish language I will recommend the Via Teldok publication (see literature).

LITERATURE

Buddine, L. and Young, B.
The Brady Guide to CD-ROM
Brady, New York, 1987, 476p.

Pettersson, R
Optiska medier
Via Teldok, no. 11, 1988, 29p.

Dataware; Corporate Guide to Optical Publishing
Dataware Technologies Inc., 1989, 39p.

Herther, N. K.
DVI is moving closer to the marketplace
Online, vol. 13, no. 2, 1989, p. 107-109

Gnislandi, P. and Campana, A.
In touch with XA. Some considerations on earlier experiences of CD-ROM XA production
Online information 89. 13th Online information Meeting Proceedings, London, UK 12-14 Dec. 1989, p. 211-226.

Pring, I.
Multi-media in the making
Computer Education, no. 62, 1989, p. 19-21

Silber, J. R.
Implementation og DVI technology for high safety training
Proceedings. Seventh Conference. Interactive Instruction Delivery, Orlando, FL, USA 22-24 Feb. 1989, p. 13-14.

Bastiaens, G. A. J.
Compact disc interactive: a multimedia system for entertainment, education and information in the nineties
Computer Education, no. 64, Feb. 1990, p. 2-5

Levine, P.
The optical outlook (optical storage)
DEC professional vol. 9, no. 1, 1990, p. 58, 60-61, 63-64, 66

Anon
Philips, Sony & Microsoft joint development of bridge between CD-ROM & CD-I
Optical Data Systems, vol. 3, no. 11, 1988, p. 3-4

Computer-Assisted Publishing with Its Standards and Compatibility Problems

Holger Wendt, Christine Zürn
Springer-Verlag, Tiergartenstr. 17, D-6900 Heidelberg

Introduction

The original task of a publishing house, that is, to publish information in printed form, has changed enormously during the past few years. The possibilities offered by electronic data processing have several effects: on the one hand, more and more authors are choosing to write their "manuscripts" on word processing systems and send them to the publishing house on data carriers or by electronic mail. On the other hand, the possibility of using one text for several purposes has increasingly led to information being published on new media.

Computer-Assisted Publishing

The term "computer-assisted publishing" is often used for publication of something written by means of electronic media (especially computers), i.e., for the application of computer-assisted methods which make it possible to find, write, design, and store information and to distribute this information via different exchange systems.

In our opinion it is important to make a distinction between the production of information by means of electronic media in combination with optical and non-electronic techniques on the one hand (e.g. keyboarding of the data, administration of the data in a database, computer-aided typesetting and printout on an electronic printer, which are mainly the task of the author) and the electronic publishing of information on the other hand, which merely means that a text is published and distributed with the aid of electronic media. We therefore tend to use the term *computer-assisted* publishing where computers are used for facilitating production and processing of printed media and *electronic* publishing where computers and communication systems serve to make data available through electronic channels.

Using New Media

From a very early stage, Springer-Verlag was concerned with the possibilities of offering data and products on non-print media. As early as in 1986, the first four demonstration CD-ROMs were presented at the Frankfurt book fair and first marketable products were produced in the following years (dangerous goods CD and mathematics abstracts CD). According to our definition, this kind of publication is part of electronic publishing and will therefore not be discussed in more detail. In this connection, we just want to stress that for good data retrieval on a CD-ROM (and of course on other databases as well) documents must be available in structured form.

Electronic Manuscripts

A publisher should, of course, aim at directly using data keyboarded by the author for further processing. Nevertheless, the problems arising are sometimes so serious that data must be keyboarded a second time.

The first problem is – although nowadays, we can possibly already say "was" – the different diskette formats, which are estimated at approximately 1300. For some time now there have been conversion systems on the market (e.g. GICO from GEFA) with which almost all diskette formats can be converted, so that physical transformation of data no longer seems to be a real problem.

The second problem, which is much more complicated to solve, is still the variety of coding systems used by the authors. Of course, many of them can be made readable by means of translation programs but this very often requires an expenditure equal to or even geater than the money saved.

In the last few years, trends have become visible in the coding systems favored by authors of Springer-Verlag: particularly in the fields of physics, mathematics and computer science, an increasing number of authors use TeX for writing and formatting manuscripts, a program which is especially designed for difficult scientific texts with many formulas. This has led us to develop so-called macro packages to facilitate author's work, e.g. with regard to layout. The use of these macros has the advantage that the expenses for correction are reduced as additional typesetting can be omitted and last but not least books and journals can be published faster. Authors in the field of medicine or biology increasingly seem to trust in Microsoft WORD. Nevertheless, the number of programs used remains tremendously high.

Document Interchange

Despite the large number of word processing systems, a publisher should aim at making a smooth-running document exchange possible. Otherwise, data cannot be used for

simultaneous use on different media. A precondition for this multiple use of data is that they are structured logically and do not contain layout-specific details. TeX or WORD, the programs mentioned above, are therefore a suitable exchange format only to a very limited extent since both of them contain mainly layout specifications.

SGML

For about ten years now, there have been serious attempts at standardization. SGML (Standard Generalized Markup Language) was developed in 1985 as a neutral exchange format for documents, and in 1986 it was made into an international standard (ISO 8879).

SGML is a tool for defining structures for documents specific to a particular application, i.e., it allows the structure of documents to be defined solely in terms of their logical structure, independent of their subsequent layout. SGML, however, does not yet offer definitions of document types. These have still to be developed by means of the SGML standard.

The first organization to take steps towards defining document types was the AAP (Association of American Publishers) in the USA. In 1986, the AAP developed an application based on SGML which became an ANSI standard. In Germany, too, efforts were made to develop SGML applications for the German-speaking countries. In 1987, the Bundesverband Druck (German Printer's Association), the Börsenverein (Association of the German Book Trade) and several publishing houses, among them Springer-Verlag, and companies in the printing industry started to work on StrukTEXT. On the one hand, printing houses were not equipped with the means of converting data, so this SGML application was only used in individual cases, and on the other hand, only relatively few documents were structured using SGML, so the necessary conversion programmes were never developed. In short – everybody waited for everybody else, and thus widespread and successful use of SGML did not occur.

Ultimate success only became possible with CALS. CALS (Computer-Aided Acquisition and Logistics Support) is an initiative of the American Department of Defense (DoD) which aims at providing a standardized coding, structuring, and exchange format for all of the DoD's documents. SGML is being used in this project for structuring the documents. When the DoD announced that as of September 1988 all calls for tenders and tenders are to be processed using this system, soft- and hardware producers reacted promptly: about four weeks after the decision, the first products supporting SGML were put on the market.

Advantages of Using SGML

When Springer publications are structured in this neutral document exchange format they can be easily and rationally processed and reused. This means that flexibility in placing typesetting and printing orders is increased. If a particular printer or typesetter is not available, it is easy to have a new edition produced by another company at short notice, as the data are coded in a way which everybody can understand and convert. A second and even more important point is the possibility of multiple use of the data. To have an article appear in more than one journal does not mean that more keyboarding will be necessary, even though these journals may have different layouts and may be produced by different printers, because with SGML only logical and therefore layout-independent document structures are defined. To offer this article in media other than paper or in a data base is then, of course, the next step to take. To proceed in this way saves on costs and – in some cases even more important – on time. That the documents have a good structure also leads to the possibility of data retrieval by content-related criteria, and new products like printing on demand, electronic profile services or hypertext applications only become conceivable with this sort of structure.

How to Proceed

First of all, one or several document type definitions (DTDs) are necessary. In this definition, the structure and meaning of the elements of a text are fixed. When SGML was developed, great importance was attached to the fact that not only computers but also humans should be able to interpret documents marked up with SGML. One should pay attention to this intention when creating a DTD.

After completion of this work it will be necessary to have individual layout descriptions for the journal, book or for any other medium which serves as the information carrier. This layout description can – or rather must – be produced by each individual supplier of information since at this point it is no longer smooth exchange of data which is in question but the layout. And layout remains individual.

The document type definition and the layout description will later be freely available to authors and keyboarders. In this way, it will be possible for them to produce their documents in conformity with SGML while, at the same time, the layout description will enable them to get a visual impression, on their monitors, of how the document will subsequently appear in the relevant journal.

SGML Activities of Springer-Verlag

If each user were to develop his own DTD, documents produced using these DTDs should then theoretically also be usable by third parties. However, in each case it would be necessary to supply information about the content of the individual elements of the structure and to change the data conversion program. From the economic point of view this is certainly not useful. If one takes into account that authors often write their manuscripts without knowing in which

journal their article will be published or even with which publisher, it becomes obvious that a DTD is necessary which is not specific to any one publisher. For this reason, at the end of 1989 we set up a working group aimed at developing and testing such a DTD. This working group consists of the publishers Springer, Elsevier, Kluwer and Thieme, the typesetters Stürtz and SRZ Berlin, the software supplier and consultant MID and the data base host FIZ.

A first practical application of SGML is the new edition of a technical handbook, the "Dubbel - Taschenbuch für den Maschinenbau". With this book we proceed as follows:

As a first step, a document type definition was made at Springer-Verlag on the basis of SGML. Then, a layout description for the printed version was developed so that a typesetter could keyboard the data. As soon as the typesetter has finished his work Springer-Verlag will receive the structured data. These data, which are then structured independently of a later layout, can easily be used both for the printed version and for other applications (e.g. on CD-ROM).

A very important aspect is, of course, the cost. Although this project is not yet finished, first calculations have indicated that expenses for typesetting have only slightly increased.

Problems Arising with the Use of SGML

Problems with SGML arise in the field of illustrations, tables and mathematics. The standard's approach for mathematics is far from being sufficient for scientific texts with many formulas. Springer-Verlag has therefore decided to use TeX, which, in the meantime, has almost become a standard for mathematical formulas. In most cases tables also can not be written by means of SGML. For figures the problem lies in the quality. There are standard exchange formats like TIFF or CCITT Group 4 Facsimile, but for high quality halftone figures none of these standards are sufficient. A further item which is still unresolved is the layout description. No functioning standards are available yet so that, at the moment, individual solutions are necessary.

Things to Do Next

The most important thing for a publisher is that authors accept the use of SGML. Most of the presently available word processing systems do not support SGML so that menu-driven text processing on the basis of a document type definition is most often not possible. Developments in this field are, once again due to the impact of CALS, foreseeable.

Further developments will happen in the field of automatically converting data produced on systems like Microsoft WORD or TeX to SGML structures and Springer-Verlag will certainly continue to contribute to these developments as well.

BROADER COMMUNICATION BANDS

by
Professor Gunnar Stette
Norwegian Institute of Technology
N-7034 Trondheim-NTH
Norway

Telecommunications is transportation of information. Before discussing the capacities of the emerging transportation systems for this type of "goods", let us look at different services and establish a set of references, both for volume of information, which is expressed in bits or Megabits, and for information transfer rates or channel capacities, which we can express in bits per second, bit/s.

We shall in this paper look at the highways for information, the broadband communications channels and systems, and we shall take this to mean systems where the transmission requires a considerable higher bandwidth or bit rate than what can be accomplished via the ordinary analogue telephone system.

Materials for broadband communication include data files, high resolution graphics (including three-dimensional and animated graphics), documents or images, and moving pictures.

Broadband communications will extend the prominent role the image now plays in newspapers, books, brochures, slides and movies, in the private, educational and business sector.[1]

INFORMATION AS AN ENGINEERING CONCEPT

It was Claude Shannon who gave us the mathematical tool to treat information quantitatively. The information contents of a symbol is related to the uncertainty of the particular information symbol, i.e. to its probability of occurence. The average information per symbol from a source which emits symbol a_n from an alphabet of N such symbols, when the probability of a_n is p_n, is given by the expression which is almost as fundamental as the famous $E = m c^2$.

$$H = -\sum_{i=1}^{N} p_i \times lg_2(p_i) \qquad (1)$$

Using the binary algorithm the information per symbol is expressed in bit (binary digit).

The required capacity of a channel for transmitting R symbols per second is therefore RxH bit/s.

The data rates associated with different types of signals differ enormously, and this is important to keep in mind when discussing the possibilities and limitations of telecommunication systems.

CONTINUOUS TRANSMISSION

The resources, such as bandwidth, transmit power, antenna area, required to establish a connection are proportional to the data rate.

For continuous transmission of information the two main types of signals are sound, in particular speech, and live pictures.

Table 1 shows data rates for CCITT standardized speech coding methods, reflecting different quality requirements[2].

CCITT Recomm.	Analogue bandwidth (kHz)	Data rate (kbit/s)
G.721	3.1	32
G.711	3.1	64
J. 42	7	192
J. 41	15	384

Table 1. Data rates for speech.

Quality	Data rate (Mbit/s)	Description
A	92 - 200	High definition TV
B	30 - 145	Digital component-coding signal
C	20 - 40	Digitally coded NTSC, PAL, SECAM
D	0.384 - 1.92	Reduced spatial resolution and movement portrayal.
E	0.064	Highly reduced spatial resolution and movement portrayal.

Table 2. Data rates for live pictures.

Live pictures require considerably higher channel capacities, typically by a factor of 1000.

A data rate of 64 kbit/s is sufficient, however, for limited resolution and limited movement, but the low data rate requires very advanced signal processing. On the top end 200 Mbit/s or higher is required for High Definition TV, HDTV. This is shown in Table 2.

DOCUMENT TRANSMISSION

We shall here use the term **document** in a wide sense, and it is then understood to comprise:

- text in alphacode,

- 64 kbit/s voice in the form of annotations

- rasterscanned graphics in high resolution.

Text in alphacode is moderate in data volume. This paper, except the figures, is contained in a data file of less than 1/3 Mbit, all control characters included. Spoken messages requires i higher data volume. With standard encoding the present text in alphacode requires the same storage capacity as about 5 seconds of speech.

Pictures turned into numbers, however, require vast amounts of information, as shown in Table 3. Screen images of 1280 x 1024 pixels and 24 bit of colour resolution require 1.28 x 1.024 x 24 10^6 bit, which is more than 30 Mbit per picture. With updating at a rate 25 to 30 pictures per second, this will require a channel of about 800 Mbit/s!

Data file or image	Data volume in Mbit
A4 facsimile (black/white)	1 - 4
A4 facsimile (gray)	9 - 16
A4 facsimile (colour)	30 - 60
Colour television image	4 - 6
High-definition colour television image	16 - 24
Newspaper page	200 - 600
High-resolution computer graphics	20 - 100
Relatively large data file	several 100

Table 3. The information contents of documents.

THE IMPORTANCE OF HIGH TRANSMISSION CAPACITY

For the transmission of document with a given data volume there is a simple relationship between channel capacity and transfer time, as shown in Fig. 1[3].

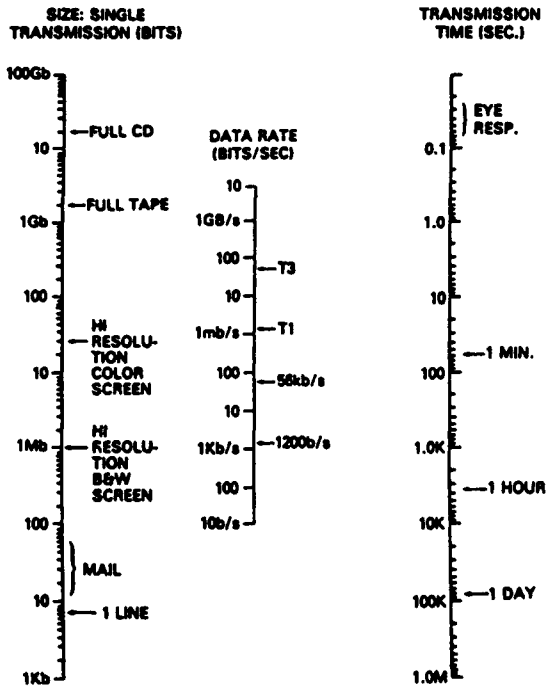


Fig. 1. Data transmission nomogram.

The present PSTN, the Public Switched Telephone Network, is designed primarily for voice, and has a very limited capacity for data transfer. Without using sophisticated signal processing methods the data rate will be in the region 2400 to 4800 bit/s.

We see now a gradual transition to what is called the IDN, the Integrated Digital Network, where speech is transmitted at the standard rate of 64 kbit/s. Access to this data stream for a user-to-user communication will then increase the capacity by a factor of 10 to 20.

From Fig. 1 it is obvious that when we go to documents with high data volumes, it is essential use channels with considerably higher capacities than what can be provided in the analogue telephgone network, or even in the IDN.

To establish usable highways for information it is necessary to have

- physical channels with sufficient capacity

- a suitable network structure that allows links with desired qualities to be established between arbitrary users.

We shall first look at the physical channels.

TECHNOLOGICAL DEVELOPMENT

The subscriber lines of the present PSTN represent a large investment, and the cost of each line is carried by a single user. It appears that subscriber lines of the twisted wire type can also be used for digital transmission with data rates up to the hundred kilobit per second range for distances of some kilometers.

Trunk routes are required to carry considerably higher data rates, but these are economically less critical since they are shared among several users. The traditional transmission systems are based on the use of coaxial cables, which provide capacities in the 100 Mbit/s range, and microwave radio relay systems, which today are constructed on a large scale with capacities of 140 Mbit/s.

Satellites have the unique feature that they cover large areas, and that they establish links between arbitrary ground stations within that area. The cost of the earth station, which is deminishing as a result of the technology development, must be carried by each subscriber, but the space segment is a common cost for the whole coverage area. Data rates are also very flexible up to the several Mbit/s range.

The real break-through for low cost data transmission is the introduction of optical fibres, which have an enormous capacity,

in the Terabit (thousand megabit) per second range. The capacity of lightwave systems has been increasing rapidly and nearly doubling itself every 18 months over the last ten years. Optical fibers of today operate at data rates of typically 45 to 560 Mbit/s, providing 600 to 80 000 voice channels per fibre. The repeater distance is in the range of 1 to 450 km.

At the same time the cost of the fibers have been decreasing. As a rule of thumb, cost per bit falls by a factor of three each time the bit rate has increased by a factor of four. The cost per meter of single-mode fiber dropped from five dollars in 1982 to twenty-five cents in 1988 and is predicted to drop to four cents by 1993[4].

The emergence of very long span high-bit-rate optical systems has reduced bit transport cost to a point where bandwidth efficiency is no longer an important parameter. This will have as a consequence that time and location loose their significance for many activities connected with information processing.

TELECOMMUNICATION NETWORK DEVELOPMENT, THE ISDN

The telecommunication networks of today form a conglomerate, because they were each constructed for a particular service. They are partly interlinked, and they make now use of many of the same physical resources, but they form separate networks.

One obvious consequence of the information theory is that a single net is sufficient for all types of services. That has lead to the idea of the ISDN, the Integrated Services Digital Network, as illustrated in Fig. 2.

ISDN represents a standardization of services in the user interface. Different types of equipment can be connected to the user interface. At the same time, within the network there will be a variety of switching and transmission functions. The ISDN is developed from the PSTN through the IDN. The transfer to the ISDN will then take place through the introduction of digital subscriber lines and more advanced signalling, that will broaden the types of services.

There are two types of user intefaces defined for the ISDN[5]:

- 2 voice channels at each 64 kbit/s and 1 data channel at 16 kbit/s, the (2 B +D) or basic rate access at 144 kbit/s for individual users. The data channel is mainly intended for signalling. Up to eight terminals can be connected to a passive bus.

- 30 voice plus 1 datachannel (30 B +D) at about 2 Mbit/s for the business user, the primary rate access. This is mainly intended for point-to-point links between PABXs (Private Automatic Branch Exchange).

In addition to the 64 kbit/s and the 16 kbit/s channels also two wideband channels have been defined, H_0 at 384 kbit/s and H_{12} at 1920 kbit/s.

The 64 kbit/s channels can be used for 3.1 kHz voice, as in the PSTN, but they can also be used for other services such as:

- data transmission

- multiplexing of several low speed data channels

- access to the data networks, both to circuit switched and the packet switched services.

- high speed (Group 4) telefax which provides higher quality and lower transmission time.

The ISDN will also ensure interconnectivity with other networks, e.g. the PTSN, to the highest possible degree, and the ISDN networks of other countries. It is anticipated that "compatible" ISDN-services in most industrialized countries, mainly in Europe, Japan and North-America, will be available from 1992.

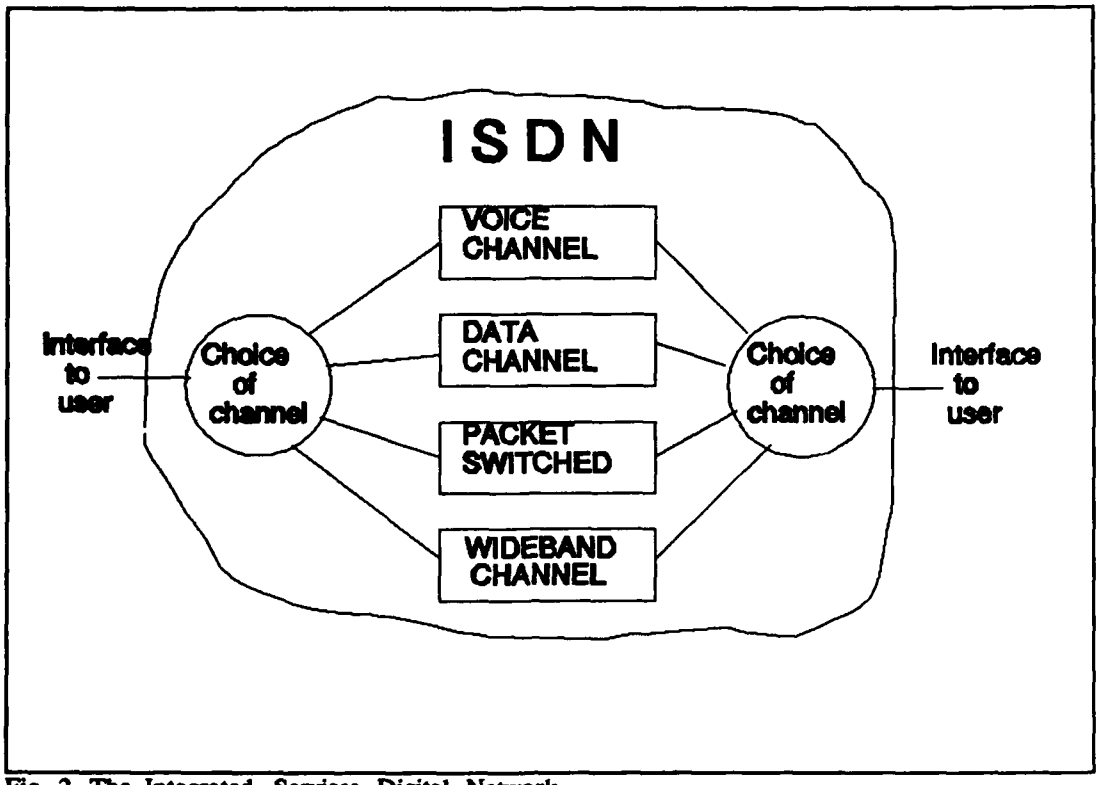


Fig. 2. The Integrated Services Digital Network.

The ISDN system as concept has some specific and important properties that are worth while emphasizing:

- The development of traditional commmunication networks is a very slow process. Therefore, the development and introduction of new services has also been a slow process. ISDN may change this situation since ISDN allows network and services with associated subscriber equipment to be developed independantly of the implementation of the network.

- ISDN will allow new services to be introduced at marginal cost, and services not demanded to be discontinued at marginal losses.

On the negative side, the developement of ISDN is partly in conflict with network optimization. You cannot develop a universal machine that is optimal for every particular application.

ISDN makes interconnection simpler, but the subscriber equipment could be more complex and costly.

It is therefore possible that viable new services developed within the ISDN may be transferred to specialized networks optimized for that particular service.

DEVELOPMENT OF WIDEBAND NETWORKS

It is important to note that the ISDN as described above can handle a wide variety of services, including live pictures, within 64 kbit/s. Nevertheless, there is a development trend towards higher transmission rates, towards the network that is usually referred to as the BISDN, the Broadband ISDN, to meet the emerging requirements for high speed document transfer.

For development of a general network the most appropriate technology is the optical

fibre, and for development of specialized networks it is satellite.

Some of the work towards the definition of the BISDN has been done within the framework of SONET, the Standard Optical Network, which defines standard optical signals, a synchronous frame structure for the multiplexed digital traffic, and operations procedures[6].

The standard was initiated in the US, and was extended to become an international standard through the CCITT (Comite Consultative de Telegraphie et Telephonie) where the first recommendation was adopted in June 1988.

The levels of the SONET Signal Hierarchy are given in Table 4.

Level	Line Rate (Mbit/s)
OC-1	51.84
OC-3	155.52
OC-9	466.56
OC-12	622.08
OC-18	933.12
OC-24	1244.16
OC-36	1866.24
OC-48	2488.32

Table 4, SONET hierarchi and data rates.

When the BISDN is concerned the capabilities must be flexible. Interfaces of about 150 to 600 Mbit/s have been proposed. The usable capacity of an interface can be organized by Synchronous Transfer Mode (STM) or by Asynchronous Transfer Mode (ATM) techniques.

STM extends narrow band ISDN concepts, as shown in Fig. 3(a)[7].

With the ATM technique all services could be carried over a single, integrated, high-speed packet-switching fabric. ATM interfaces are extremely flexible, using signalling to dynamically reconfigure a mix of logical channels, as illustrated in Fig. 3(b).

USER EQUIPMENT

Also the user equipment is undergoing a development which means that the new capabilities of the wideband networks can be utilized.

Today, a typical PC operates at 2 to 8 MIPS (million instructions per second), contains a few megabytes of memory and offers "effective" back-plane bus and I/O rates of a few megabytes per second.

Emerging PCs and workstations will operate at 20 to 100 MIPS, have extensive memory capacity, and I/O rates near 100 megabytes per second.

High end machines operate at 500 MFLOPS (million floating point operations per second) to 2 GFLOPS, have up to 2GB of memory, and channels in the 100 Mbit/s range.

Machines under development will operate at 10 to 40 GFLOPS, have upwards of 8GBit of memory, and channel interfaces upwards of 800 Mbit/s.[8]

SPECIALIZED NETWORKS.

Satellites may be used as a part of the general telecommunication network, but they may also be used to established overlay networks, - user to user, independent of the general telecommunications systems. This type is called user oriented satellite systems. For a satellite system, the high capacity pipelines can be redirected instantaneously over the whole coverage area through the signalling and access control system.

The America company AT&T initiated in May 1986 a socalled VSAT (Very Small Aperture Terminal) system providing business customers with digital data and video communications.

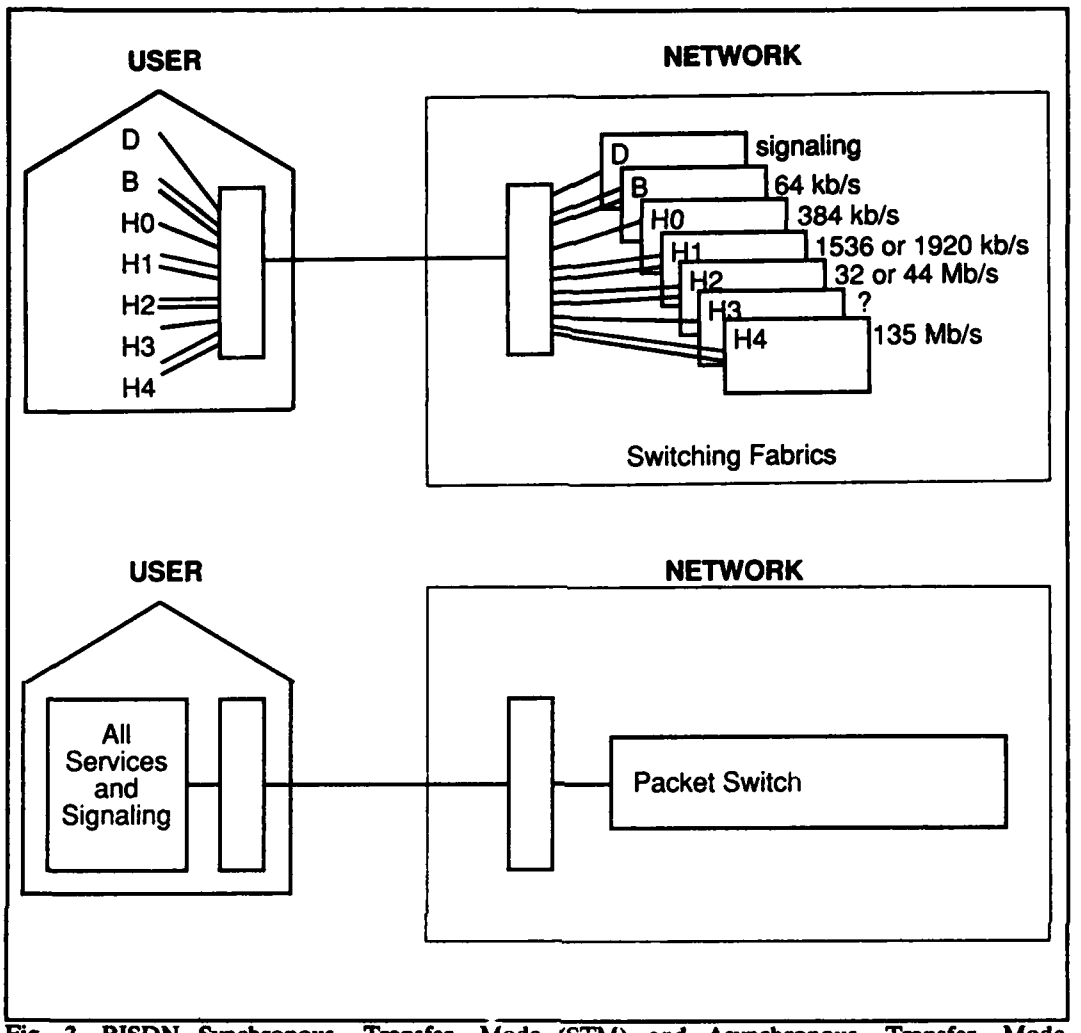


Fig. 3. BISDN Synchronous Transfer Mode (STM) and Asynchronous Transfer Mode (ATM).

One-way transmission from the hub-station is offered at data rates 9.6 kbit/s, 56 kbit/s and 1.5 Mbit/s. Two-way transmission is offered at 256 kbit/s outbound from the hub and 56 kbit/s inbound.

One version of a VSAT system, which has been developed by ESA (the European Space Agency) is the APOLLO-system, shown in Fig.3. The system is a very good illustration of its basic ideas.

One characteristic feature of the APOLLO system is communication links with different capacity in the different directions. The forward direction is a time shared 2 Mbit/s link from the information sources via a large earth station to small and cheap receiving stations at the users' premises. The returns could be provided via terrestrial networks or via a low data rate satellite link.

Another development of the VSAT concept is the SWAN (Satellite Wide Area Network), as illustrated in Fig. 5[9], which is used to interconnect LAN (Local Area Network) and MAN (Metropolitan Area Networks).

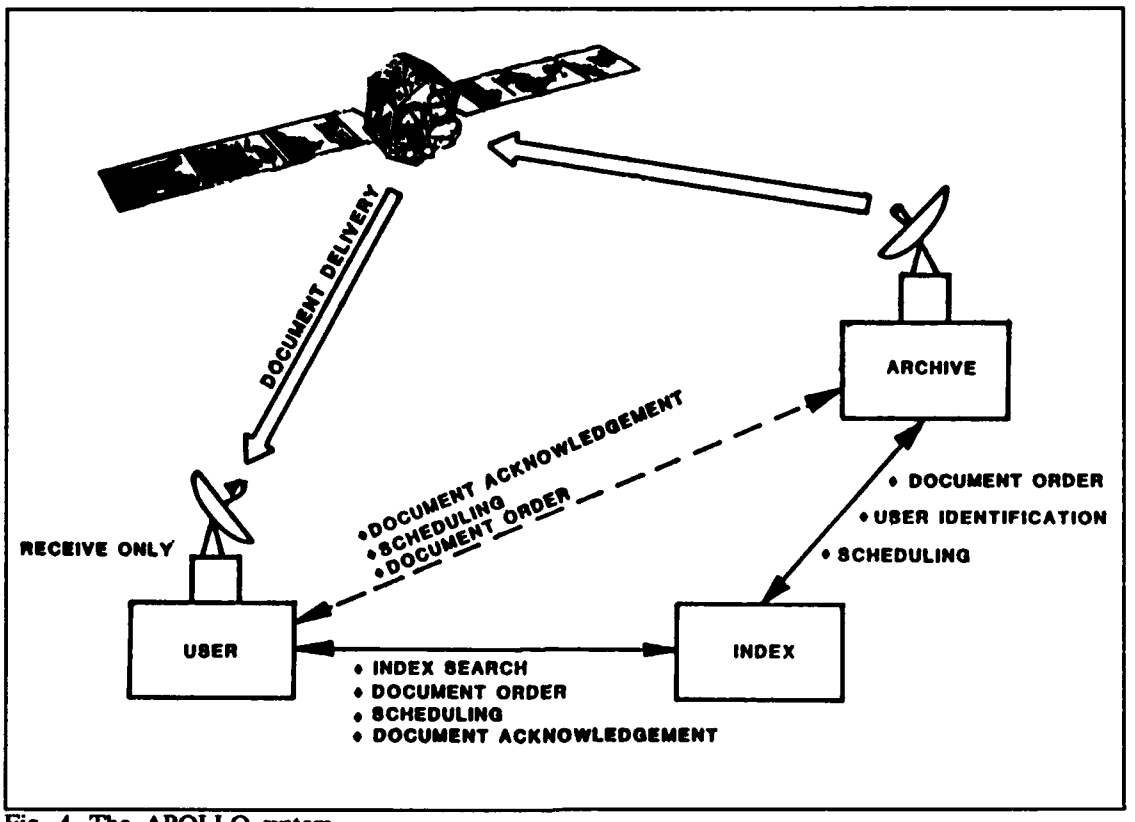


Fig. 4. The APOLLO system.

This is a network with a very flexible architecture, and can be used in different forms of point-to-point, star and mesh configurations.

One particular category of applications for this type of system is wide area communication services and remote area information services of different types such as corporate data base access and public/private information database access.

POINT TO MULTIPOINT SYSTEMS

Another category of specialized networks can be established via broadcasting satellites, which represent both enormous investments and very valuable commercial possibilities.

The cost of the small stations is largely a question ov production volume. We have recently seen satellite TV receivers sold in shops with a price tag of £199, everything included.

A new signal format, MAC (Multiplexing of Analog Components), allocates about 10 % of the total TV channel capacity for sound and/or data on an interchangeable basis. And 10 % of a TV channel is a considerable capacity when used for voice and data!

The D-MAC system, which is now gaining the most widespread acceptance, can provide about 2 Mbit/s of data in addition to ordinary TV program sound. (If the TV system is used for several simultaneous multi-lingual commentary channels, the capacity is reduced accordingly.) The data can be directed to defined user groups by the access control system, which initially was deviced to document the numbers of viewers of a particular TV-program, and to ensure that the viewers pay for the service. The organization of the access control is shown on Fig. 6[10].

The MAC system with access control, and with encryption as required, is a powerful infrastructure for distributing documents

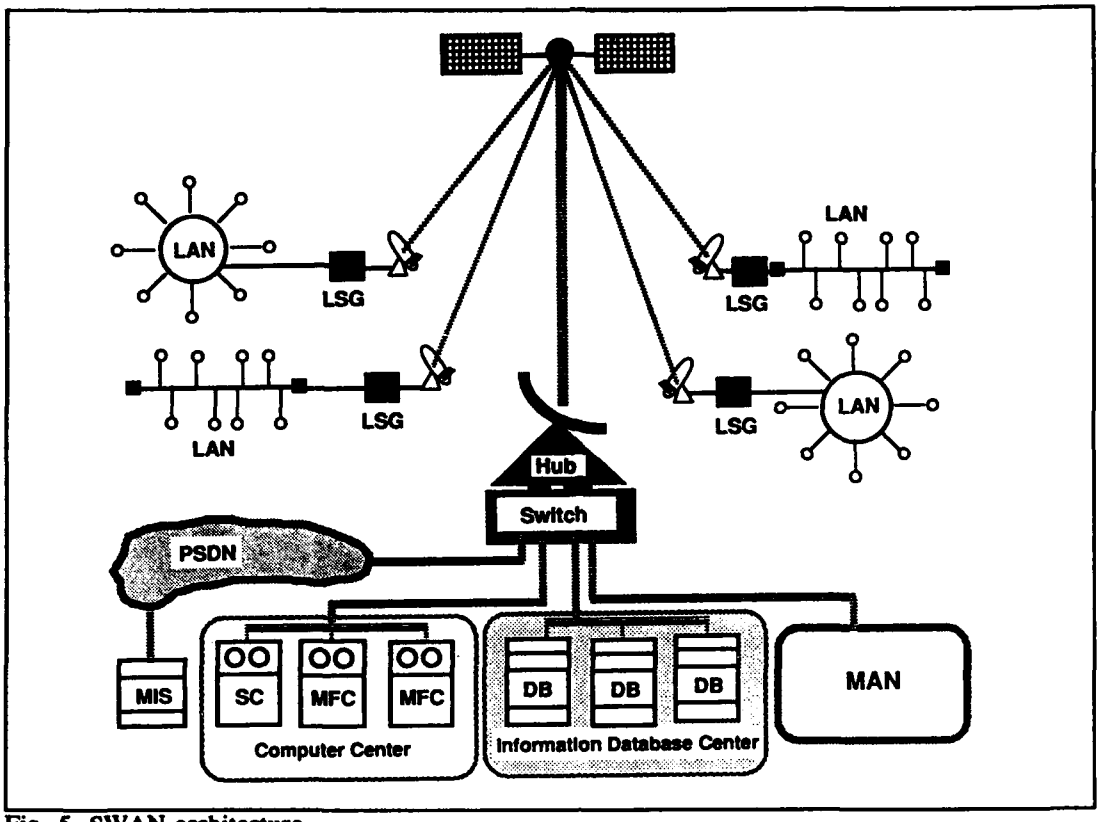


Fig. 5. SWAN architecture.

and other types of data. The system can also be utlized for closed TV-transmission, e.g. to branch offices of geographically widespread companies.

Planning of sessions, ordering of documents, etc. can be arrranged via the ordinary terrestrial network. The biggest advantage of this system is low equipment cost. The configuration could be a commercial TV satellite receiver, an interface card and a commercial PC.

CONCLUDING REMARKS

This paper has aimed at giving an overview of the development of information transportation systems with capacities beyond what is available today via the public switched telephone network.

Two elements are directing the course of development, technology and business, and the latter is the more difficult one. We are today able to provide technical solutions to almost any problem. The real challenge is to develop services that are of value to the user, and that the user is willing to pay for.

Service development is a difficult process, since the value of a service in many cases depends on the penetration of its use. We have also the situation that new services require new equipment, which will be developed and mass produced only if there is a service creating sufficient demand. This circle is not easy to break.

The ISDN, which allows a new service to be introduced at marginal cost, is an important tool for testing out the value of further service to the users.

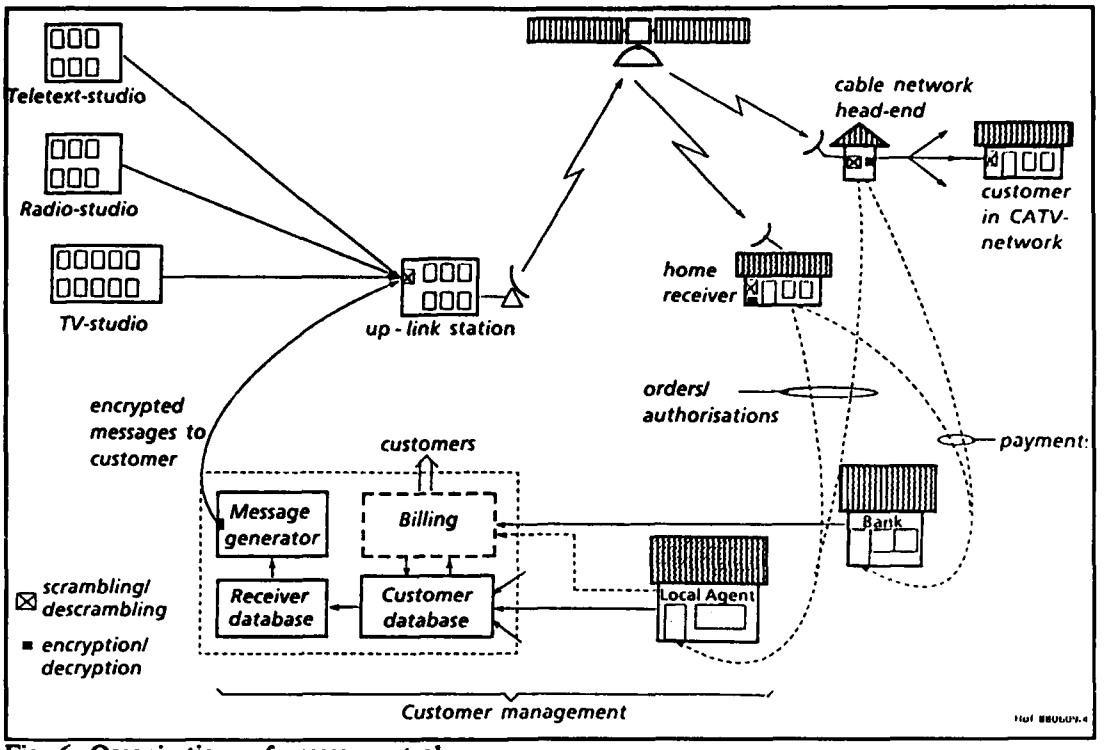


Fig. 6. Organization of access control.

REFERENCES

1. Armbruster, Heinrich, and Arndft, Gerhard: Broadband Communication and its Realizetion with Broadband ISDN. IEEE Communications Magazine, Vol 25, No 11, November 1987.

2. Minzer, Steven E., and Spears, Dan R.: New Directions in Signaling for Broadband ISDN. IEEE Communications Magazine, February 1989.

3. Roberts, Michael M.: ISDN in University Networks. IEEE Communications Magazine, Vol.25, No 12, Dec. 1987.

4. "Further Gains for Japan Predicted in Telecom War". Lightwave, September 1986.

5. Netland, Bjørn: Datakommunikasjon i ISDN. Overføringstjenester og anvendelser. TELEKTRONIKK Nr 3, 1989.

6. Ballart, Ralph, and Ching Yau-Chau: SONET: Now it's the Standard Optical Network. IEEE Communications Magazine, March 1989.

7. Minzer, Steven E. and Spears, Dan R.: New Directions in Signaling for Broadband ISDN. IEEE Communications Magazine, Feb. 1989.

8. Chlamtac, Imrich, and Franta, William R.: Rationale, Directions, and issues surrounding High Speed networks. Proc. of the IEEE, Vol.78, No.1, January 1990.

9. Sundara Murthy, K.M., and Gordon, Kenneth G.: VSAT Networking Concepts and New Applications Development. IEEE Communications Society Magazine, May 1989.

10. Fjøsne, Hans: Unpublished memorandum.

Fields of Application of the Processing of Natural Languages with the help of Artificial Intelligence

by

Bente Maegaard
EUROTRA-DK
University of Copenhagen
Njalsgade 80
DK-2300 Copenhagen S
Denmark

1. Introduction

Natural Language Processing (NLP) is almost as old as computers. The first applications of computers for manipulating linguistic data were those that were the most easy to make: statistics, sorting (establishing word lists from text), concordances etc. In brief, the results were interesting for researchers as they provided new data on language and linguistics, and much faster than humans could do. Some results were however readily usable also by people outside the researchers' world: all those having to do with coding and decoding messages. As you all know, computers were used for this purpose already during World War II. I will not go further into this field of application.

Since the late 40'ies and the early 50'ies many fields of application have developed, supported by progress in computational linguistic research and in production of computers, and lately also by progress in the field of Artificial Intelligence (AI). A real breakthrough of the industrial application of NLP has however not yet taken place. Foreseeing this breakthrough some people have created a new term Language Industries, or in French where it was first invented "l'industrie des langues".

We can give no exact date for when this industry will become profitable, but it seems obvious that it will happen in the 90'ies.

In this paper I have chosen to treat a number of fields of application, each of them by at least one case example. The paper does not postulate to cover all fields of application, nor to cover any application exhaustively, - for obvious reasons.

Terminological remark:
The scientific community does not have one single opinion on the relationship between NLP and AI. NLP people seem to believe that there is a field called NLP which exists independent of AI, and which can take over methods and results from AI. Most AI people however seem to think that AI is a wide field covering e.g. NLP. The wording of the title of this paper shows that the latter view is not followed here. Still this does not solve the problem of defining AI. In the present paper a rather relaxed interpretation of the field is assumed, e.g. that one deals with conceptual structures or objects, that some kind of implication or inferencing is used, or similar. In order to get examples from a variety of NLP fields even this relaxed view is not always adhered totally to.

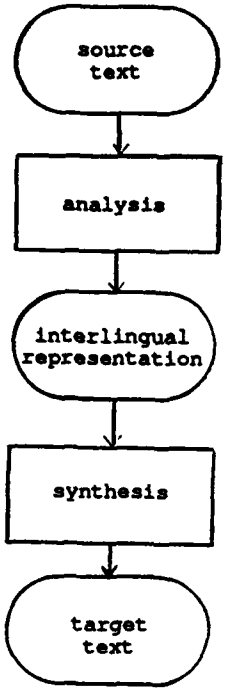
2. Machine Translation (MT)

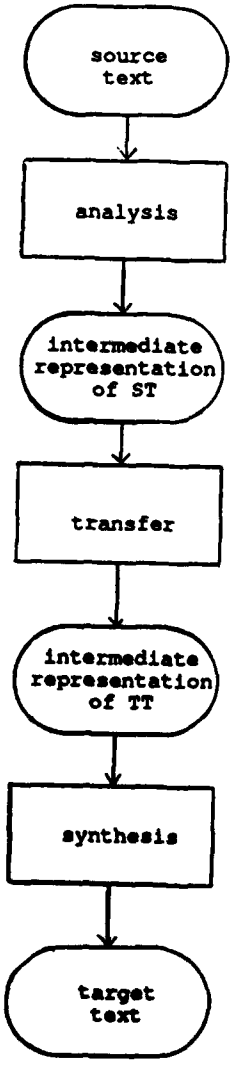
A major application field for NLP is the field of translation. It is also one of the very first fields of application: the Georgetown system which translated Russian - English was presented in 1954, and a variety of systems are being marketed at present. Translation is however a very complex task, and at present no general-purpose MT system exists which produces good quality translation. This will continue to be true for some time. The reason that MT is anyway one of the major applications, is that there is such a need for translation that a less than ideal output can still be useful. In order to give a state-of-the-art picture of the field of MT, I have chosen to describe two MT projects, a European and an American one, and to mention a few more.

2.1. The task of translating by computer

In the early days of machine translation the translation task was actually seen as a variant of the decoding tasks mentioned above. Secondly, there was a feeling that translation could more or less be reduced to consulting a bilingual dictionary, and making a few (local) word order transformations.

The difficulties observed with these systems led to the development of second generation MT projects where the translation process was broken down into either two or three steps:





The difference between these two approaches is that the former assumes an interlingua, i.e. a (formal) language which has the expressive power of any natural language (or at least those involved), whereas the latter is also modular, but maintains a translation phase proper, the transfer phase. There are no examples of general translation systems of significant size using the interlingua approach, one reason being that a useful interlingua has not been definable, but recently a side version of this has emerged: the DLT in which the artificial (but not formal) language esperanto is used as interlingua. The original idea of an interlingua - if at all feasible - would undoubtedly make the use of AI necessary.

The transfer approach has been used widely, with emphasis on various aspects.

2.2. EUROTRA

The Council of the European Communities decided in 1982 to start a European (EEC) project in MT, covering all the EEC official working languages, financed jointly by the Commission of the European Communities and the member states. The project as such is ending at the end of this year, but follow-up actions are being prepared.

The novel aspect of EUROTRA, when it was first conceived is the fact that it is inherently multilingual where other systems which can treat several language pairs (e.g. SYSTRAN and METAL) normally develop each language pair more or less separately.

EUROTRA is transfer based. The result of the analysis phase is an internal representation of the source text. In the transfer phase, the internal representation of the source text is translated into an equivalent internal representation of the target text. This translation step basically consists in exchanging lexical units, but in some cases a change of grammatical relations is also necessary. Finally, the synthesis phase produces a surface target language text from the internal representation.

The advantage of this modular structure of the overall translation process, is that the same analysis module can be used for all eight target languages. Similarly, the same synthesis module is used for all source languages, leaving to the transfer module only to be bilingual.

In a transfer system which is multilingual there is a clear interest in obtaining as small transfer modules as possible, i.e. in pushing as much of the work as possible into the monolingual modules. This means that the research will aim at producing internal representations for exchange which are as interlingual as possible, thereby eliminating the need for explicit transfer rules which change an object (a lexical item or a feature or a structure). The greater part of the linguistic research done within EUROTRA is aimed at improving the interlinguality of the internal representations.

All the pieces of knowledge which are used, are essentially of linguistic nature. In the follow-up programme it will be investi-

gated how extra-linguistic knowledge can also be used for e.g. disambiguation purposes.

2.3. KBMT-89

At Carnegie Mellon University the KBMT-89 project ran from 1988 to 1989. This machine translation project translates between English an Japanese (in both directions). It has two major characteristics: it is knowledge based and it uses the interlingua approach to MT.

The interlingua is a structure, describing the semantic relations between constituents of the text. The meaning of the words of the text are given by an ontology or domain model which contains 1500 concepts.

Apart from the traditional syntactic parser and syntax to semantics mapping, and apart from the (static) knowledge expressed in the ontology, the KBMT contains an episodic or dynamic memory which establishes e.g. instantiation links between objects or events.

Following the traditional interlingua idea source and target language expressions are never compared in KBMT, the concepts and semantic relations being the common part.

As described above the interlingua approach is quite ambitious, and it will be extremely interesting to see extensions of the KBMT approach to a larger coverage, so it is a project to be followed. It should be mentioned as conclusion that the KBMT-89 is a research project, not aiming at a practical MT system. Future projects may also include application.

2.4. Interlingua or transfer?

This discussion is still ongoing among MT researchers. It seems to me that the most clear advantage of the interlingual approach is that it makes the tedious transfer writing superfluous. And the most clear disadvantage is that it is at present not possible to define an interlingua for more than very restricted sublanguages. The KBMT-89 researchers agree with these two statements, but claim further that transfer MT requires post-editing while interlingual MT does not. In my opinion all MT, if treating more than extremely restricted sublanguages, will require post-editing and this for some time to come.

While is seems pretty obvious that it is a goal of transfer systems to make the transfer component as small as possible, it is still not clear that the optimal strategy is to make the transfer component disappear totally. Also the fact that knowledge bases and inference techniques are used does not automatically entail the interlingual approach. Still, a good deal of research into the theory of machine translation and into the way the various approaches relate to such a theory is needed. In fact this subject is attracting interest by researchers in the field.

3. Computer Assisted Instruction

Another field of application of NLP is Computer Assisted Instruction (CAI). Computers have been seen as possible tools for instruction for more than 20 years. The early CAI programs would teach students e.g. fortran, by presenting statements and questions to the student. The answers would be checked against the answers stored and various actions taken: the student would get feedback and the teaching program would continue with the next question or go back and repeat some early information and questions belonging to this. I.e. each program would itself contain the strategy for the teaching expressed in the control structure of the programming language, and the data (questions and answers) in some language.

The next step was the creation of socalled author languages for writing teaching programs (e.g. COURSEWRITER). This would free the author from expressing the control of the teaching event in a programming language like pascal or fortran, which was a big step forward. On the other hand what has almost always hampered the development and use of these CAI systems has been the poor treatment of natural languageand author language have normally not offered much support in this area.

The problem is that all text to appear either as instruction, as question or as answer has been stored in a fixed format. For the computer part, the instructions and the questions, this is not so much of a problem. The problem arises with the treatment of the human part, the answers. If there is no linguistic treatment of answers at all, the program has to be written in a way that only very concrete and short answers are needed, e.g. "What is the capital of Norway?" "Oslo", or "How many countries are there in the EEC?" "12". This puts a limit to what subjects can be treated by CAI. Even in this very restricted scenario, however, authors of programs realise that some relaxation or intelligence is needed: if the program is not a spelling exercise, some deviation from normal spelling should be accepted, both capital and small letters, and typographical errors, and the choice between "12" and "twelwe" etc. There has been some extensions to this type of answer handling, but still real full scale NLP using parsers and grammars for natural language is not used. The reason for this is undoubtedly that the requirements of robustness and correctness of the grammars involved can hardly be met. Since it is a claim of Language Industry that such grammars with a reduced scope and function can now be made, this may change in the near future. On the other hand such intelligent answer-handlers will only be of real use when they are not only working correctly, but also efficient in terms of computer time and space.

When we see these years a growing interest in CAI, it is not because of the integration of large-scale NLP components. Rather the revival of this application stems from the fact that 1) in many countries or areas of countries it is the best way of spreading knowledge, 2) computers are becoming cheaper, 3) not all subject fields need linguistic expression equally much and by ingenious use of various techniques good interfaces can be obtained. This leaves a challenge for NLP people.

3.1. Expert systems and CAI

Some research has been carried out to investigate the use of expert systems or expert system technology for CAI purposes. This is probably the most obvious way of including AI techniques.

An expert system contains knowledge about a given domain, and inferencing rules which express the various types of implications between pieces of knowledge. An expert system for teaching needs to contain all the knowledge about the domain, but also knowledge about how to correct errors. An experimental teaching system CAPRA for teaching introductory programming concepts to novice programmers has been developed in Spain (Garijo et al. 1989). The system builds a program to solve a specific problem, following a programming methodology. It is able to explain the reasoning process step by step, showing the knowledge used at every decision point. This is the expert system part of the CAI program. This part is combined with a module that supervises the interactive process between the student and the expert system.

The supervision module uses a natural language interface for communication with the user; it explains problems and proposes exercises, depending on the student's performance.

The research prototype works but still there is some way to go before this methodology will be widely applied. One should also note that introductory programming concepts is one of the easier fields to describe exhaustively in a knowledge base, and that it will definitely take some time before complex domains can be described adequately, just like for expert systems in general.
Research projects as the one described are interesting and seem promising for the integration of expert systems into CAI.

An interesting point put forward by expert system researchers is the fact that there is actually a difference between expert systems and domain components of intelligent CAI systems. The difference being that the reactions and implications to be made by the expert system depends on whether the user is an expert or a student. This means that the factual knowledge can be common to the two kinds of use, whereas the procedural or strategic knowledge will differ according to the use (Kamsteeg and Blerman, 1989). Consequently, this puts a constraint on how the knowledge is structured in the expert system if one wants to use it also for teaching purposes.

3.3. A challenge

One very appealing and yet unexplored idea is to view an NLP grammar and dictionary of a language as a knowledge base, and to create an intelligent CAI system based on this knowledge. It is obvious that a considerable amount of knowledge engineering is needed in order to create the knowledge base in an expert system format. The claim is that some of the formalisation effort has already been made in transforming the grammar and dictionary information from the format of grammar books and ordinary dictionaries into formal grammars and dictionaries for NLP. How important the next step, i.e. the transformation of such computer-oriented knowledge into an expert system knowledge base is, remains to be seen.

4. Application of neural networks

An approach to modeling intelligence which is now an alternative to the traditional AI approach, is the use of neural networks (Rumelhart and Clelland 1986). The idea is basically to model neural networks in a computer by defining states and transitions between states in a multi-level organisation. Such a network is then trained by giving it input and the corresponding correct behaviour. E.g. if the task is hyphenation of words, the network will be given a certain amount of words with indication of all possible hyphenation points. When given a new word the network will hyphenate with a certain, and normally rather good probability of correctness.

The hyphenation application has been developed e.g. for Danish, and it works reasonably well. On the other hand hyphenation programs made according to more traditional principles may perform as well, so as a research result this application is interesting, but seen only from the application point of view it is of less interest.

It may be argued that hyphenation is a rather simple task; it is mentioned here because it constitutes a real application of neural networks to a NLP problem. Below I will describe the application of neural networks for a subtask of NLP, namely syntactic category disambignation.

One of the major problems in NLP is ambiguity and the overgeneration this leads to. Ambiguity arises at all levels of processing, e.g. for each word of a written text there may be an ambiguity of word class (Eng. can is either a noun or a verb), within a word class there may be several readings or meanings (can as modal verb or full verb), etc. Such ambiguities may lead to several possible structures and semantic interpretations. It is one of the difficulties of NLP to disambiguate or choose between parallel structures. For efficiency reasons disambiguations should take place as early in the processing as possible. Benello et al. 1989 used a neural network to disambiguate syntactic categories.

Since people use the context to disambiguate words, the network was also given a context, 4 unambiguous words preceeding the unknown (ambigious) word, and one following. After training it determined syntactic category with 95% accuracy. This is comparable to results of wellknown linguistic parsers, - but again like for the hyphenation: not better than other known methods.

The interesting aspect of this type of research is consequently not really the already obtained application result. The interesting aspect is the new way of modeling intelligence. It is particularly interesting because there are reasons to believe that the human brain works much the

same way: first of all it is clear that human brains have strong constraints on the amount and type of computation they can do, and these constraints may weaken traditional linguistic theories claim of psychological reality. No human has the capability of capturing at the same time all 90 possible parses of "Still water runs deep" which a traditional computer program has to examine before choosing the correct one.

Apart from this issue on psychological reality, it may also be that neural networks find larger use in NLP applications. The future will show what happens when neural networks are built which have to cope with more complex problems.

5. Summary

Two very important areas of NLP applicaton have been described in some detail, machine translation and computer assisted instruction. Both fields are extremely important, and of growing importance. Artificial Intelligence techniques are only starting to be used in applications. At the same time an alternative model for artificial intelligence has emerged: neural networks. Neural networks are interesting from a theoretical point of view, because they can be said to take into account the biology of human information processing. It is not possible at present to evaluate the potential of neural networks for application in the Language Industries.

References

Benello, J., A.W. Mackie and J.A. Anderson: Syntactic category disambiguation with neural networks, Computer Speech and Language, vol. 3, no 3, Academic Press, 1989.

Campbell, J.A. and J. Cuena: Perspectives in Artificial Intelligence, vol. 1-2, Ellis Horwood Limited, Chichester 1989.

EUROTRA is described in a special issue of Machine Translation, Kluwer Academic Publishers (forthcoming), and in a special issue of Multilingua.

Garijo, F.J. et al: CAPRA: an intelligent system to teach novice programmers. In: Campbell and Cuena, vol. 2, p. 179-196.

Garijo, F.J. and F. Verdejo: Knowledge representation for teaching programming in an ICAI environment. First IEEE Conference on Artificial Intelligence Applications, Denver, Colorado, 1984.

Kamsteeg, Paul and Dirk Blerman: Differences between expert systems and domain components of intelligent tutoring systems. In: Campbell and Cuena 1989, vol. 2, p. 197-208.

KBMT-89 Project Report, Center for Machine Translation, Carnegie Mellon University, 1989.

Multilingua, 5-3, Mouton, Amsterdam, 1986 (on EUROTRA).

Rumelhart, D.E. and McClelland (eds.): Parallel distributed Processing: Explorations in the Microstructure of Cognition (vol. 1: Foundations), MIT Press, Cambridge, Massachusetts, 1986.

FULL TEXT RETRIEVAL WITH GRAPHICS

by

Michael Lesk
Bell Communications Research
Room 2A-385
435 South Street
Morristown, NJ 07960
United States

Abstract

Conventional full text retrieval systems often omit the pictures from the material they display. We are taking the existing machine-readable text of the American Chemical Society journals, scanning the pages from microfilm and extracting the images from the text by algorithms which analyze the digitized bitmaps. The combined combined pictorial and text material will then be used with full-text search to provide access to the complete file. The major experiments to be done are to: (a) measure acceptability of the electronic systems; (b) compare full text search with search of titles, abstracts and/or indexing; and (c) compare presentation of full page images in bitmap format, presentation of text in ASCII with graphics on demand from images, and traditional paper copies of the journals.

The major parts of this research are

(1) Software to partition digitized images of pages into textual, tabular and pictorial areas. This is used to extract the graphics material, which is then matched with the commands referencing the pictures in the typesetting tapes, and prepared for display as bitmaps.

(2) Search software which implements conventional searching capability (Boolean and coordinate index term search) on the full text of the journals, which is available from the typesetting and on-line operations of the American Chemical Society. Experiments are also continuing with the use of singular value decomposition to group documents and concepts to aid searching.

(3) Browsing and reading software to help people read complex journal material on-line, by highlighting matches, formatting paragraphs, and providing interactive screen displays.

Introduction.

Despite many predictions over the last thirty years that electronic information systems would entirely replace paper (see for example Samuel 1964), even in advanced societies scientific publications are still mostly read in traditional form. This is true despite the availability of most textual material in a machine-readable format, thanks to the general adoption of word processing systems. We suspect that part of the problem is the lack of graphical material in many conventional full-text retrieval systems, many users of which see words without illustrations as a very inferior substitute for traditional journal publication. Our experiments provide various formats of full text with graphics to see what, if any, kind of electronic presentation will attract users away from conventional publications.

This effort is a joint project between the American Chemical Society, Chemical Abstracts Service, OCLC, Bellcore and Cornell University. Our test area is chemistry, thanks to (a) the availability of the American Chemical Society backfile of full-text chemical journals online, (b) the availability of the Chemical Abstracts Service file of indexing and bibliographic information, and (c) the quality of, and interest in information retrieval shown by, the Cornell chemistry department as users.

In this paper we will discuss some of the questions raised by the preliminary work so far and the various systems being proposed to answer them. The major issue is how to efficiently integrate graphics with full text. People skim journal articles very rapidly, and with pictures we can not compensate for slow display by searching rapidly, since it is difficult to search pictures.

2. Information content.

Our collection is based on the text files of the American Chemical Society. These are derived from their computer typesetting production facility and contain a very detailed markup of the individual articles (for example, the sentence boundaries are marked). There are approximately 500,000 pages of material available. Our current 1,000 article database is about a 1% extract (in our counting, an "article" may include something as small as a one-paragraph book review). It represents the first twelve issues of the Journal of the American Chemical Society for 1988. Approximately 20% of the pages, measuring in square inches, are figures. There are about 4200 pages in the file, and the total text is about 30 Mbytes.

We have three sources of data that we are using to construct the file. These are (a) the ACS text files, (b) the Chemical Abstracts Services files of indexing, abstracting, and bibliographic data, and (c) the microfilm versions of the original journals. The intent is to get the text from the ACS and CAS files, and the pictures from the microfilm. The reason we are scanning microfilm rather than paper is that microfilm, being a physically robust material, can be scanned faster in bulk. A Mekel M400 microfilm scanner is used to produce images at either 200 or 300 dpi, and the pages are then processed in several steps.

The first step in processing is to sort printed from supplementary pages. The Journal of the American Chemical Society provides a facility for authors to

supplement their articles with pages of additional data which are not published in the journal as mailed, but which are on the microfilm. We have to identify and eliminate these pages in order to provide a file which matches the printed journal. Three basic techniques are used to distinguish printed text pages from other pages. The supplementary pages are most commonly prepared on ordinary typewriters. Note that all techniques depend entirely on fairly gross properties of the images, rather than any kind of optical character recognition.

(1) Average bit density. The typical text block of densely printed typesetting reaches a density of over 20% black bits (but never as much as 50% black). Any page which contains a region which achieves the typical text density is deemed to contain text.

(2) Line spacing. The spacing between lines is remarkably accurately determined by an autocorrelation computation. The number of black bits on each horizontal scan line is counted, and used to make a vector of horizontal density vs. page position. This is actually done twice, once for the left half of the page and once for the right half. The vector is then autocorrelated at each possible shift relative to itself; the first peak in the autocorrelation function turns up the line spacing. The line spacing used in JACS text is about 0.135 inches (10 points); the line spacing used in the table of contents is about 0.265 inches (19 points). Any page with the interlinear spacing characteristic of JACS typesetting is noted as a text page.

(3) Columnation. JACS is normally printed in double column. Each page image is examined for vertical strips of white space extending at least 1/3 of the way up the page (some pages have full width tables at the top). The number of such vertical strips are counted and used to decide how many columns this page used; most of the supplementary pages are multi-column tables. Any page with two columns of about the right width is again marked as a text page.

It would seem that with three different ways of marking pages as text pages, this algorithm would be failsafe: that is, it would be more likely to mark supplementary pages as printed pages than the other way around. Unfortunately the errors are about balanced (although rare: some dozens of failures in 7000 images). The reason is that some printed pages are filled or almost filled with one large table or figure, and thus do not have the characteristics of a densely printed text page. The mistakes made in the classification are caught as the page numbers are matched up with the images.

The next step is reassembling certain page images which have been split in the microfilming. When an article ends in the middle of a page, and has supplementary pages after it, there is a microfilm image for the portion of the page which contains the end of the first article, followed by images for the supplementary pages, followed by a microfilm image for the portion of the page which contains the beginning of the next article. To obtain the appearance of the journal as printed, we logically OR together the bits from the two text halves, having deleted the supplementary images in the previous step.

Finally, some hand examination is required per journal issue. Not only do we spot check the previous steps, but each issue contains a few pages of prefatory material (e.g. instructions to authors) which may look like text but are not part of the paginated sequence of the main journal.

We now must match up the images of the articles with the text of the articles. The text tapes contain authors, titles, and so on but not page numbers. However, the Chemical Abstracts Services tapes do contain the page numbers, and can be used to make the necessary table of article titles vs. page number. For issues where, for administrative reasons, we are processing the issues before the Chemical Abstracts tapes have arrived, it is not difficult to type in the list of page numbers for each article.

At this stage we have data files which contain the full text of each article, including figure captions and tables, in ASCII, from the American Chemical Society data. We also have files containing the abstract, index terms, and bibliographic citation from the Chemical Abstracts data. The full article is also available as page images.

The next step is to identify the specific figures which correspond to each figure referenced on the text. There are four kinds of graphics which we identify (tables, figures, equations and 'schemes' – chemical structures) using techniques similar to those above for classifying pages. These will be discussed in detail in a later paper; for previous work on this subject see Fletcher and Kasturi (1987). For each picture of table, the typesetting tapes specify the size, but not the position on the page. However, the sequence of figures within the article is given, and this is used to match up the recognized images with the text references. Similarly, the footnotes and the Ascii forms of the tables are matched up with the places where they are reference (this, of course, does not involve any image processing).

3. Presentation.

There are several different ways in which the users wish to access information in these journals. We distinguish at the moment three major forms of access which require quite different facilities:

(1) Reading a particular article whose citation they have.

(2) Looking for information on a particular subject.

(3) Skimming through issues for current awareness or general intellectual curiousity.

In order to make the machine-readable form acceptable, we have to serve each of the information needs above. It is not clear that a single computer interface can provide for all of these purposes. Note that for purpose (2), searching for a particular subject, the printed journal needs to be supplemented also, in this case with the printed version of Chemical Abstracts. In addition to questions of presentation and style of interfaces, we must also decide what kinds of informa-

tion are presented to the user and in what format. The biggest such question is how to use the Chemical Abstracts indexing effectively since it contains both textual structure and term normalization, facilities that users could exploit to improve their searches but which many users may not understand well enough to use without some aids.

Interviews with the users have found that skimming is a major part of their use. In this mode, they flip through rapidly, and claim that they rely mainly on the pictures. Chemists, particularly organic chemists, are very skilled at interpreting chemical structure diagrams, and visualizing three-dimensional structures of the compounds they study. Such visualization is very important in their work, whether it be synthesis of molecules or analysis of their properties. Thus, rapid display of properly drawn pictures is important; neither systematic nomenclature nor line notations nor bad approximations of the actual structures will suffice.

For these reasons, the American Chemical Society suggested that we keep and display the page images as well as the text plus pictures separately. This will permit us to make some kind of measurement of the value of the original typography vs. a reformatted text. We support at present three different interfaces for detailed physical presentation of the material and expect to see which of these the chemists prefer. Each one deals in a different way with inability of current computer workstations, even large screen workstations, to equal the resolution of printed paper. Typical screen resolutions are 72 or at most 100 dpi, and a maximum of perhaps 1000 vertical pixels down a page; printing is at least 1000 pixels per inch, and even conventional laser printers and copiers have several times the accuracy of the typical screen.

These different interfaces will include Superbook, the OCLC Diadem system, and a simplistic system called Pixlook. The Diadem system will be described in a later publication; the other two are presented here.

(1) Superbook, by Egan, Gomez and Remde (1989), displays Ascii text with graphics on demand. The advantages of displaying Ascii are that the display is better matched to the capabilities of the computer terminal, it is possible to reformat the text to match the window size chosen by the user, and that the displayed text can be matched to the users' query. For example, after a word or phrase search, the text display begins with the start of the paragraph in which the users' search terms appear. The terms are highlighted. In addition, another window displays the tables of contents of the material so that the user can locate the current page in the context of the overall publication. Searching in Superbook is based on co-occurrence of terms within paragraphs; a variety of aids such as term truncation and aliasing are maintained as well.

(2) Pixlook displays images of the original pages. It performs coordinate index or Boolean searches, and is able to search specific document fields and also supports truncation and suffixing. For each search, the matched list of titles is presented and the user can choose which articles to read. Each article is presented first as a display of an image at 100 dpi resolution, 1 bit per pixle, from the original page; this means that nearly the entire page can fit in a window on a workstation with a 1000x1000 display. The user can then zoom in on a portion of the image to 200 dpi resolution, which is normally adequate for reading (100 dpi is quite adequate for a quick view, and can be read, but is not satisfactory visual quality for most readers for the long term). The user can move around in the image or move backwards and forwards in the text. It is also possible to browse the pages of tables of contents as images and not use the search features at all.

A variant on this program, for those who wish to see images primarily, displays only the pictures from the articles which match the search terms. These can be brought onto the screen and examined quickly in low resolution. Then, the user can pop up the full pages, with text, from the articles whose pictures seem interesting.

As examples Figure 1 shows a sample of a Superbook screen, while Figure 2 shows a Pixlook display at 100 dpi and Figure 3 at 200 dpi. We are planning to run experiments on the comparative acceptability to the users of these different presentation formats. Experiments with Superbook in the past have shown superior performance for searches aimed at specific target information. Not only is the searching efficient, there being no feasible way to match with paper the facilities of full-text electronic search, but the display is effective at calling the user's attention to the material found. However, we have not evaluated Superbook formally in applications such as skimming, nor for the problem of known-item retrieval in a large document collection. Image-based displays may well be superior in these applications since they retain the format the users find familiar and which has been tailored over the years for effective use by chemists.

4. Searching.

Not only do the different systems we have implemented include different searching methodologies, but the user will have a variety of sources of information to search. There are of course the usual facilities of title, author, abstract and the full text terms. In addition Chemical Abstracts defines for each article a very precise set of phrases, for example *phenyloxirane, prepn. of, by alkene epoxidn. in presence of nickel catalysts* which are printed in their subject index (in this case alphabetically under Oxirane). These entries are defined for the purpose of display in an index people are browsing, whether on paper or electronically. Thus, we can provide the user a choice of whether to search for free text, selected index terms, or author names or other entry points; and also of several different interfaces and ways of specifying the searches.

The obvious danger, particularly in such a large collection, is that people will be drowned in retrieved items. Whether this is based upon text retrieval (words like 'NMR' or 'bond' appear more than once per article, on average), or upon figures (there are an average of six graphic items per article), there is a danger of over-

loading the user. For different problems, there will be different strategies of dealing with the amount of information available. Some of these strategies depend upon the user and some upon the computer.

In the Pixlook program, for example, there are basically only conventional facilities for narrowing searches. The intent is to provide very fast pop-up of matched pages and rely on the user to recognize what he is interested in. In the right circumstances this can be very fast (for example, in the process of checking the page images to see whether the pictures had been correctly noted 50 bpi images were used, and although these can not be read, they can be reviewed at better than one per second). We would like to compare the attractiveness of fast review of images to adding Boolean ANDS to queries as a way of dealing with a large number of retrieved documents. Note that in this interface the user must type the search terms and there is no facility to help suggest any.

By contrast Superbook provides a variety of facilities for structuring the retrieval to give the user a view of what has come back. The overall list of articles comes with indications of which sections of the journal contain how many instances of each term, and thus permits the user to feel in some way oriented. As the user expands the table of contents, more detailed indications of how many terms occurred in which sections are shown. Thus, the user can maintain some context and feeling for how the retrieved passages are distributed around the documents. This works best, of course, in a collection which contains one continuous and organized document, but still has value in the context of a collection of many articles, for each has subsections. Superbook also makes it easy to select terms from existing material to be searched again, using the mouse; the user can thus not only be prompted by looking at the document but also need not worry about spelling the terms correctly.

We have also briefly investigated clustered methods of searching. The original intent was to see if the articles could be grouped, instead of by issue, by computing term overlaps and then running hierarchical clustering algorithms. However, these clusters did not seem consistently intuitively sensible to some chemists who considered them, and so we are going to use the sections of Chemical Abstracts into which the articles are classified. This divides the collection into fewer parts than would be possible with automatic clustering, but the categories make sense to chemists and are relatively familiar.

A more interesting question is whether any clustering can be done of the graphical figures. The importance of diagrams to the chemists would make it attractive to have some technique for grouping the figures. Since we only have bitmaps, it would be difficult to do this directly, but Tom Landauer at Bellcore has suggested that we could classify the figures on the basis of the text of their captions. This is not a complete answer; many of the structural diagrams are printed as 'schemes' and do not have captions, for example. But it would be very useful to be able to search for figures which are similar to other figures or to present them in some content-related way.

5. Conclusion.

Many people insist that they will never abandon paper for any kind of screen display. They point to the amount of reading they do removed from workstations and networks (on airplanes, in traffic jams, even on canoe trips). Electronic information delivery, however, makes possible searching and browsing in ways that potentially offer greatly improved performance, if we can manage to deliver this in a way the users find appropriate, acceptable and effective.

Our system supports a variety of approaches to retrieving both text and images for use in chemical information delivery. Preliminary interviews with chemists and experience with these systems suggests that effective presentation of graphics is very important for user satisfaction with the system. We are planning a series of experiments to measure the utility of

- a variety of kinds of information to search, including both free-text, fielded data, and indexing;
- a variety of search strategies, including browsing indexes, Boolean and coordinate indexing search, and other techniques;
- a variety of display strategies, including images of original pages and resynthesized text.

We expect to learn what kind of computer facilities a system must include in order to cope with the wide variety of user information needs that chemical information satisfies today, and if possible to improve our ability to respond to these needs in a way the users like.

Acknowledgments.

This project is a collaboration with Lorrin Garson of the American Chemical Society, Martha Lindeman of OCLC, Jim Lundeen of Chemical Abstracts Service, and Jan Olsen of Cornell University. The assistance of Karen Bogucz is gratefully acknowledged.

References

[Egan] "Behavioral Evaluation and Analysis Of a Hypertext Browser," D. E. Egan, J. R. Remde, T. K. Landauer, C. C. Lochbaum, and L. M. Gomez, pp. 205-210 in *Proceedings of CHI'89 Human Factors in Computing Systems*, Austin, Texas, April 30 – May 4, 1989.

[Fletcher] "Segmentation of Binary Images into Text Strings and Graphics," L. A. Fletcher and R. Kasturi, *Proc. SPIE Conf. on Applications of Artificial Intelligence V*, vol 786, pp. 533-540 (1987).

[Samuel] "The banishment of paperwork," A. L. Samuel, *New Scientist*, vol. 21, no. 380, pp. 529-530 (27 February 1964).

Note

In response to requests, some of the viewgraphs used during the presentation are included here (pages 5-8 to 5-13).

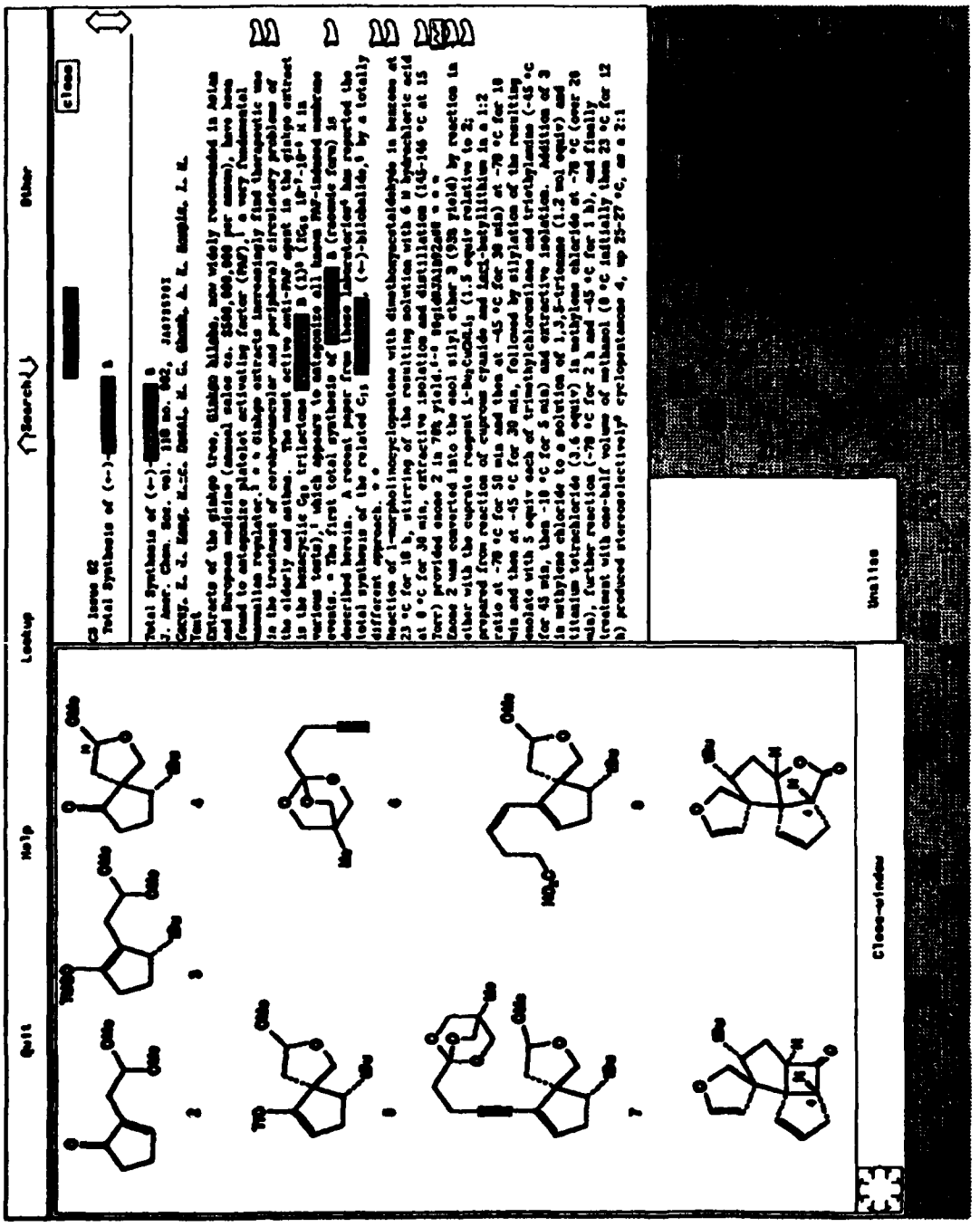
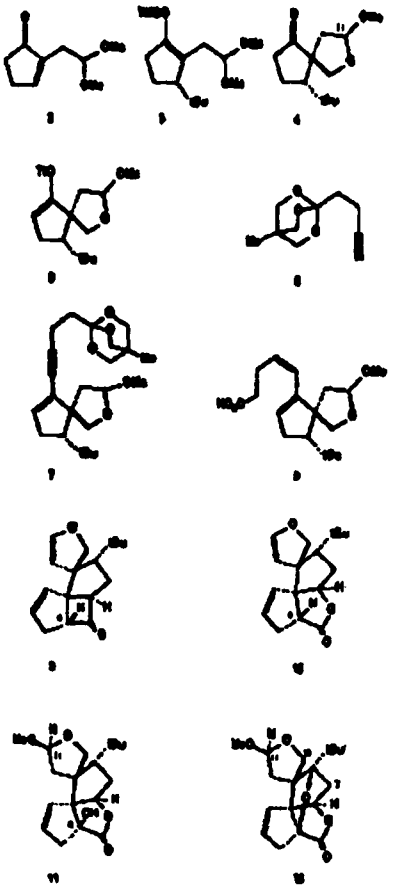


Figure 1. Superbook example.

5-6

Figure 2. 100 dpi sample.



isolation and silica gel chromatography (SGC) an 80% yield of enol triflate 5. A solution of 5 and $Pd(PPh_3)_4$ (5.7 mol%) in benzene was stirred at 16 °C for 15 min and then treated successively with a benzene solution of acetylenic OBO ester 6[11] (1 equiv), *n*-propylamine (2.3 equiv), and 0.5 equiv of cuprous iodide, all at 16 °C to give after 4 h at 16 °C, extractive isolation and SGC 76–84% of the coupling product 7 (2:1 mixture of anomers), mp 44–47 °C.[12] The triple bond of 7 was reduced by reaction with 1.5 equiv of dicyclohexylborane in tetrahydrofuran (THF) (0 °C for 2 h, 23 °C for 0.5 h), followed by protonolysis (acetic acid 23 °C for 16 h), and decomposition of residual boranes (H_2O_2, 23 °C, pH 10). The resulting solution was acidified to pH 3 with 1 N hydrochloric acid, brought to pH 11 (vigorous stirring, 4 h) and reacidified to pH 3 to cleave the OBO ester unit,[13] and the (*Z*)-olefinic acid 8 was isolated by extraction and removal of solvent (70% yield, colorless oil). Conversion of 8 to the corresponding acid chloride (3 equiv of oxalyl chloride in benzene at 23 °C for 2 h) and addition of the acid chloride in toluene solution (0.2 M) over 2 h to a stirred solution of tri-*n*-butylamine (10 equiv) in toluene at reflux, followed by further reaction at reflux for 1 h, furnished stereospecifically (71–87% yield) the tetracyclic lactone 9, mp 59 °C. Structure 9, which results from internal ketene-olefin cycloaddition[14] and elimination of the anomeric methoxy group (under tri-*n*-butylammonium chloride catalysis), follows from spectroscopic data and the transformation 11→ 12 described below.[15]

Addition of 9 in acetone (at −30 °C) to a stirred solution of triphenylmethyl hydroperoxide in 8:1 acetone–1 N aqueous sodium hydroxide at −30 °C over 10 min and a further reaction time of 2 h at −30 °C produced a single Baeyer–Villiger product 10, mp 163 °C, in 86% yield.[15] Lactone 10 was transformed into 4-hydroxylactone 11 (ginkgolide numbering as in 1) by a two-step sequence: (1) deprotonation (1.5 equiv of sodium bis(trimethylsilyl)amide in THF at −50 °C for 20 min) followed by reaction of the resulting anion with 2 equiv of (*E*)-2-(phenylsulfonyl)-3-phenyloxaziridine[16] (at −50 °C for 5 min and then at −50 °C to 0 °C over 10 min) to afford the corresponding α-hydroxylactone (73% after SGC) and (2) exposure to a 1% solution of camphorsulfonic acid (CSA) in methanol at 23 °C for 48 h to give 11, mp 155 °C (75%).[17] Reaction of 11 with lead tetraacetate (4.5 equiv) and iodine (3 equiv) in pyridine-1,2-dichloroethane at 5 °C under sunlamp irradiation for 10 min resulted in complete conversion to a single product, determined by 500-MHz ^{1}H NMR analysis to be the cyclic ether 12,[18] rather than the hoped for product of functionalization at C(12). Although this result was not useful as a synthetic step, it did provide confirmation of the stereochemistry of intermediates 9, 10, and 11.

The required oxygen bridge between C(4) and C(12) was established by an alternative route starting from 10. Reaction of 10 with 1.2 equiv of propane-1,3-dithiol and excess titanium tetrachloride in methylene chloride at 0 °C for 10 min and then at 23 °C for 40 min produced the thioacetal–primary alcohol 13, mp 230 °C (90%), which was transformed into the aldehyde 14, mp 165–166 °C (75% yield), by treatment with pyridinium dichromate (PDC, 1 mol equiv), powdered 3-Å molecular sieves and acetic acid in methylene chloride at 0 °C for 1 h. The aldehyde 14 was converted into the bis-acetal 15 (80% overall yield as a 2:1 mixture of C(12) anomers) (major anomer from SGC, mp 107 °C) by the following process: (1) oxidative dithiane cleavage by reaction of 14 with 0.5 mol equiv of periodic acid in 1:1 methanol–methylene chloride containing ca. 1% water at −30 °C initially then at 0 °C for 20 min and 23 °C for 40 min and (2) stirring of the resulting product with methanolic CSA at 23 °C. The C(4)–C(12) oxygen bridge was generated by the following sequence: (1) deprotonation of bis-acetal 15 with use of 1.9 equiv of lithium diethylamide initially at −25 °C and then at 0 °C for 15 min and subsequent oxygenation with (*E*)-2-(phe-

(10) McMurry, J. E.; Scott, W. J. *Tetrahedron Lett.* **1983**, *24*, 979–982.

(11) The preparation of 6 was carried out as follows. 4-Pentynoic acid was converted to the acid chloride (oxalyl chloride, benzene, 23 °C) which was treated with 3-methyl-3-hydroxymethyloxetane to form the corresponding ester (85% overall) and in turn rearranged with boron trifluoride in methylene chloride at −78 °C to form 6 (92%). See: Corey, E. J.; Raju, N. *Tetrahedron Lett.* **1983**, *24*, 5571–5574. 4-Pentynoic acid was obtained in 25% overall yield by the following sequence: (1) alkylation of diethyl malonate in ethanol with propargyl chloride; (2) saponification with potassium hydroxide in aqueous ethanol at 23 °C; and (3) thermal decarboxylation of propargyl malonic acid at 150–170 °C over 45 min.

(12) (a) Yamaguchi, K.; Tohda, Y.; Hagihara, N. *Tetrahedron Lett.* **1978**, 4667–4671. (b) Ramachandran, V.; Liotta, D.; Eickhoff, G. *Synth. Commun.* **1980**, *11*, 917–923.

(13) Corey, E. J.; Ha, D. *J. Am. Chem. Soc.* **1984**, *106*, 2735–2736.

(14) See: (a) Corey, E. J.; Desai, M. C.; Engler, T. A. *J. Am. Chem. Soc.* **1985**, *107*, 4339–4343. (b) Corey, E. J.; Desai, M. C. *Tetrahedron Lett.* **1985**, *26*, 3535–3538. The stereospecificity of the internal cycloaddition to form 9 was predicted from mechanistic considerations[14b] and the latter degree of steric screening for the pathway leading to 9.

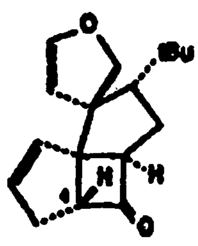
(15) The structure of 10 was confirmed by the conversion of 11 → 12. Use of *tert*-butyl hydroperoxide as oxidant or higher reaction temperatures led to the appearance of the positional isomeric lactone as a byproduct.

(16) (a) Davis, F. A.; Stringer, O. D. *J. Org. Chem.* **1982**, *47*, 1774–1775. (b) Davis, F. A.; Vishwakarma, L. C.; Billmers, J. M.; Finn, J. *J. Org. Chem.* **1984**, *49*, 3241–3243.

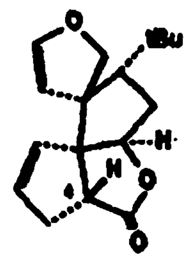
(17) Methyl acetal lactone 11 was obtained as a single kinetically controlled stereoisomer (from 500-MHz ^{1}H NMR analysis). The orientation of methoxy at C(11) follows from the strong steric shielding by *tert*-butyl at the opposite face of the tetrahydrofuran subunit. The orientation of the hydroxyl group at C(4) follows from the strong preference for formation of a cis 3,5-fusion in the hydroxylation reaction and is further confirmed by the conversion to 12.

(18) The ^{1}H NMR spectrum of 12 (with spin decoupling) provides unambiguous support for this structure and the following key assignments: H_{10a}, δ 4.33 ft, $J_{10a,11}$ = 9.6 Hz; H_{10b}, δ 3.45 ft, $J_{10b,10a}$ = 9.6 Hz, H_{10b}, 4d, 2.60 δ, $J_{10b,10a}$ = 13.1 Hz, $J_{10b,11}$ = 7.7 Hz; H_9 d, 1.98 δ, $J_{9,10b}$ = 13.1 Hz; *t*-Bu, s, 1.07 δ. In all intermediates in the synthesis which have H attached to C(9) a coupling $J_{9,10}$ of 5–6 Hz is observed, the doublet for H_{10} then shows there is no H attached to C(9).

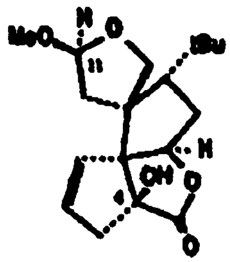
Figure 3. 200 dpi sample.



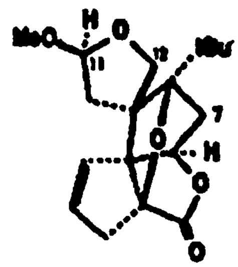
9



10



11



12

isolation and silica gel chromatography (SGC) an 80% yield of enol triflate 5. A solution of 5 and $Pd(PPh_3)_4$ (5.7 mol%) in benzene was stirred at 16 °C for 15 min and then treated successively with a benzene solution of acetylenic OBO ester 6[11] (1 equiv), *n*-propylamine (2.3 equiv), and 0.5 equiv of cuprous iodide, all at 16 °C to give after 4 h at 16 °C, extractive isolation and SGC 76–84% of the coupling product 7 (2:1 mixture of anomers), mp 44–47 °C.[12] The triple bond of 7 was reduced by reaction with 1.5 equiv of dicyclohexylborane in tetrahydrofuran (THF) (0 °C for 2 h, 23 °C for 0.5 h), followed by protonolysis (acetic acid 23 °C for 16 h), and decomposition of residual boranes (H_2O_2, 23 °C, pH 10). The resulting solution was acidified to pH 3 with 1 N hydrochloric acid, brought to pH 11 (vigorous stirring, 4 h) and reacidified to pH 3 to cleave the OBO ester unit,[13] and the (*Z*)-olefinic acid 8 was isolated by extraction and removal of solvent (70% yield, colorless oil). Conversion of 8 to the corre-

---

(10) McMurray, J. E.; Scott, W. J. *Tetrahedron Lett.* 1983, *24*, 979–983.

(11) The preparation of 6 was carried out as follows. 4-Pentynoic acid was converted to the acid chloride (oxalyl chloride, benzene, 23 °C) which was

Page 658 (j moves down, k up, l right, h left, z zoom, q ends)

Why not?

"Libraries for books will have ceased to exist in the more advanced countries except for a few which will be preserved at museums."

– Arthur Samuel (IBM)

prediction of 1984, made in 1964

Remember microfilm?

Don't these quotes seem familiar?

Microfilm promises "to have an impact on the intellectual world comparable with that of the invention of printing" – 1936.

Microphotography is "one of the most important developments in the transmission of the printed word since Gutenberg" – 1940.

Not only did hypertext not invent text, it didn't even invent hype.

Conclusion:

Just fast page turning isn't enough:
need searching to make a difference.

The Problem

People search online, but they don't read online

Why?

- Costs too much
- Bad interfaces
- No graphics

Needed for Solution

Extraction of graphics
Procedures for display
Different human interfaces

Our Plan

Create an automated chemistry library

Text: from the American Chemical Society file
  10 years, 20 journals, key U.S. publisher
  full text with typographic markup
Pictures: scanned from microfilm
  graphics extracted from page images
  most graphics are line drawings, e.g.
    chemical structures
    spectra
    reaction pathways
  about 20% of journal is pictures (sq. inches)
Interface
  Free text searching or index term searching
  Page image display or ascii display of text
  Browsing interfaces as well as searching
Experiments
  Electronics vs. paper
  Different electronic modes
Design of a better system for chemists

Handling Images

1. Identifying text
   (microfilm contains 40% supplementary pages)
   – overall density must be 20% black
   – line spacing 10 points (autocorrelation function)
   – columnation (look for white strips)

2. Matching text to articles
   – need table of contents

3. Skew removal
   – mostly we lucked out

4. Identifying figures, tables, schemes and equations
   above parameters plus
   – aspect ratio
   – horizontal and vertical lines
   – presence or absence of captions
   – white space distribution

Interfaces

Superbook
- table of contents, fisheye
- text with marginal icons
- free text word search
- popup graphics

Alternatives being provided
- Boolean search
- representative icons for graphics
- Page image display
- Different resolutions

Experiments

Display: Paper vs. Images vs. Ascii

Search: Boolean vs. Coordinate Index

Vocabulary: Free text vs. Indexing

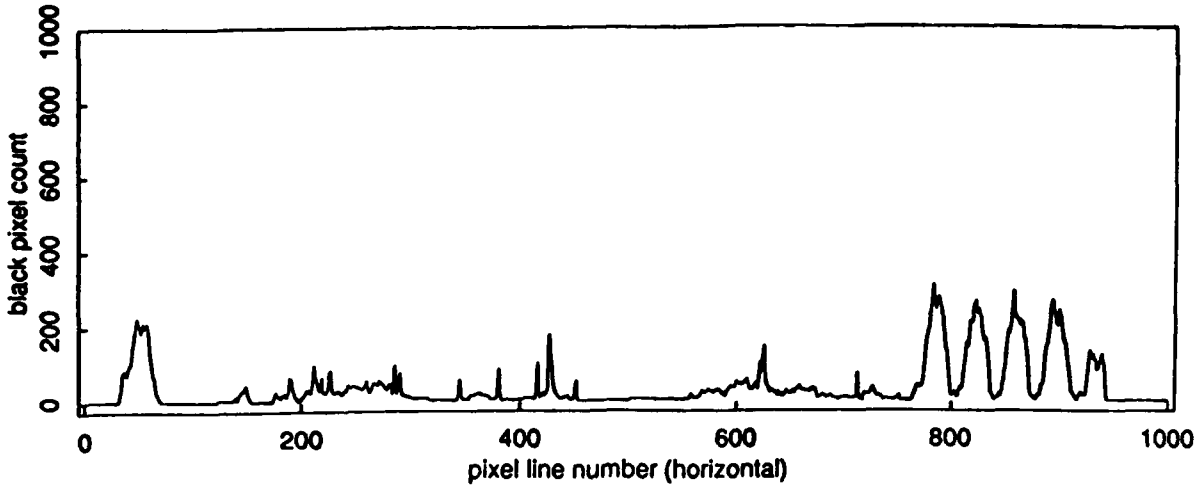
Tasks: Known item vs. Subject vs. Current Awareness

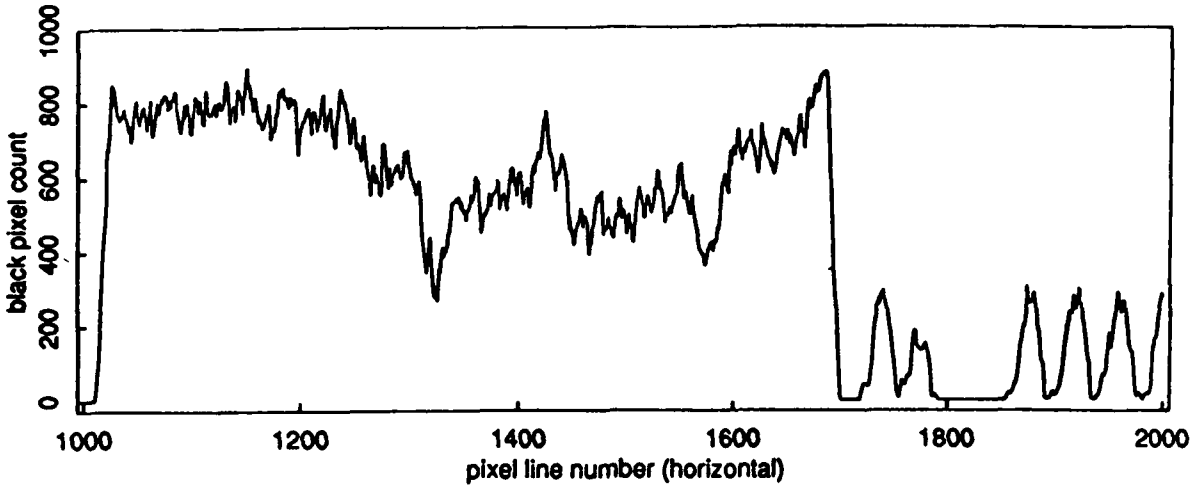
Conclusions

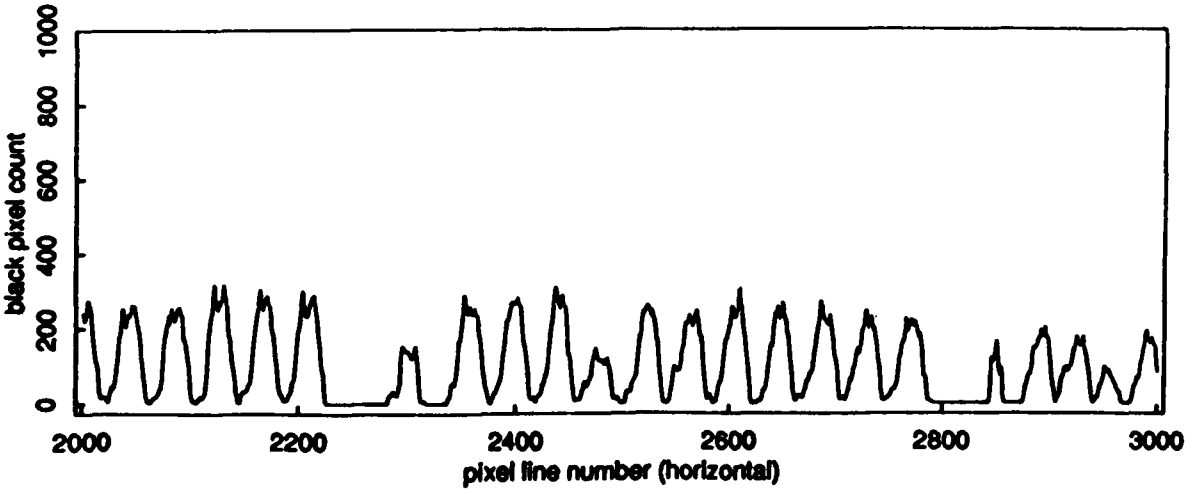
This is harder than we expected!

Image technology permits
- domain independent view
- retention of typographic quality

I believe we will have a system the chemists will prefer to paper.







Horizontal Projection Profile for Page 7162, Column 1

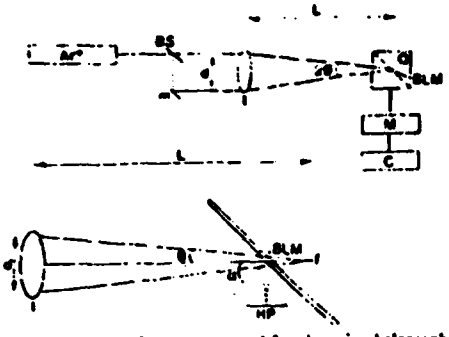


Figure 1. Schematics of the system used for observing holographic interferometry in BLMs. Ar^+ = argon ion laser; BS = beam splitter; m = mirror; Q = quartz cuvette; BLM = bilayer lipid membrane; M = microscope; C = camera; HP = holographic plate; f = focus; I = lens; and d = interbeam distance.

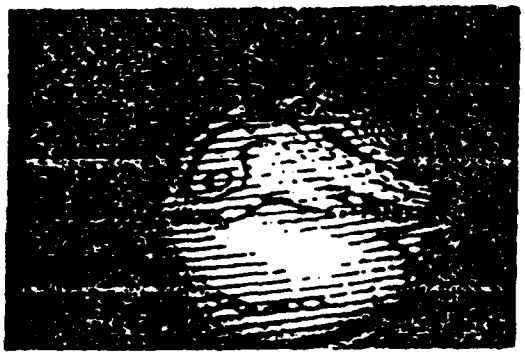


Figure 2. Complex interferometric fringe pattern observed in a thick GMO film prior to thinning to a BLM.

Second, advantage has been taken of fluorescence to produce distinct interference patterns of molecules that are incorporated onto the surface of a BLM. The power of holographic interferometry is demonstrated in the present report by showing differences between thick lipid films and true BLMs, as well as by determining the shapes and sizes of cadmium sulfide (CdS) particles in situ generated on glyceryl monooleate (GMO) BLMs. Evidence is also provided here, by holographic interferometry, for the presence of Merocyanine 540 on the surface of GMO BLMs.

Experimental Section

Merocyanine 540 (Sigma), glyceryl monooleate (GMO, Nucheck Co.), cadmium chloride (Aldrich), and hydrogen sulfide (Matheson) were used as received. Water was purified by means of a Millipore Milli-Q system.

BLMs were formed across a 1.00-mm-diameter hole in a thin (0.10–0.15 mm thick) Teflon film, placed diagonally in a rectangular quartz cuvette. The cell was filled with 2.0 mL of water at ambient temperature. BLMs were made, as reported previously,[5-11] by "painting" decane solutions of freshly prepared (ca. 200 mg in 1.0 mL) GMO across the pinhole. Thinning of the initially formed film to a black, bimolecular (50 ± 5 Å), thick lipid membrane was monitored by observing the light reflected by the BLM. Illumination was provided by a 150-W xenon lamp via a 500-nm cutoff filter and an optical filter. The reflected light was observed through an Olympus PM-10-M microscope coupled to a TV monitor and videorecorder via a NEC NC-8 CCD camera. Interferometric patterns were photographed with a 35-mm Nikon M-35S camera using Kodak color print film (400 ISO). Exposure times were typically 0.1 s.

Merocyanine 540 was introduced into one side of the solution bathing the BLM by injection (30 μL, 10^{-3} M aqueous stock). Semiconductor formation on the BLM has been described.[7] Briefly, subsequent to establishing the presence of a true BLM, 200 μL of 10^{-3} M $CdCl_2$ was introduced into one side and a stoichiometric amount of H_2S was slowly injected into the opposite side of the BLM. Within 5 min, the semiconductor particles became visible and grew, over the next 1–2 h, to a thin film on the BLM. Semiconductor-containing BLMs proved to be highly stable and rigid.[7]

The schematics of the experimental setup used for the holography are shown in Figure 1. The beam of the argon ion laser (Spectra Physics 2020, 100–200 mW) was divided by a beam splitter (BS). One half of the beam was allowed to impinge directly on the BLM. The other half was diverted by a mirror (m) prior to reaching the BLM. This arrangement produced a fringe pattern on the BLM.

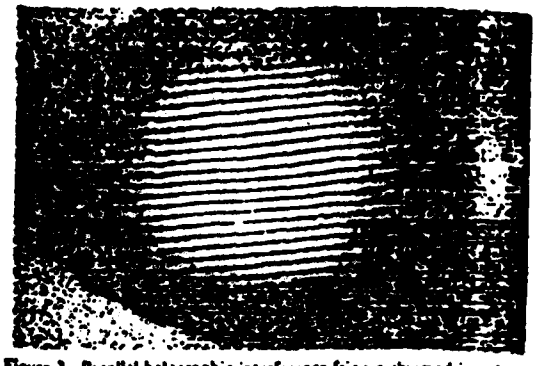


Figure 3. Parallel holographic interference fringes observed in a [illegible] 2-nm-thick GMO BLM.

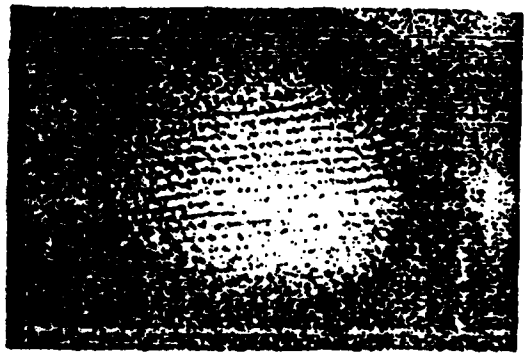


Figure 4. Interferometric fringe observed in 50 ± 5 Å thick GMO BLM subsequent to the in situ deposition of CdS particles.

Results and Discussion

A typical, complex interferometric-fringe pattern observed during the thinning of the GMO is shown in Figure 2. The parallel lines are indicative of flat sections, while the contours are attributable to thick local patches of the surfactant. With thinning, the contours move around and gradually diminish. Appear-

(7) Baral, S.; Zhao, X. K.; Rotondi, R., Fendler, J. H. *J. Phys. Chem.* 1987, *91*, 2701. Zhao, X. K., Baral, S.; Rotondi, R.; Fendler, J. H. *J. Am. Chem. Soc.* 1988, *110*, 1012.
(8) Zhao, X. K.; Fendler, J. H. *J. Phys. Chem.* 1986, *90*, 3446.
(9) Zhao, X. K.; Fendler, J. H. *J. Phys. Chem.* 1988, *92*, 3350.
(10) Zhao, X. K.; Hetrick, P. J.; Fendler, J. H. *J. Phys. Chem.*, in press.
(11) Baral, S.; Fendler, J. H. *J. Am. Chem. Soc.*, in press.

```
Please type query or ? for help
heme protein
Article Matches    Title
     39       2 Heme Imidazole H-Bonding
     86       2 Recombination of Myoglobin with CO, O2, NO, and Isocyanides
     94       2 Electron-Transfer Kinetics of Zn-Cyt c
     96       2 Zinc/Ruthenium-Modified Myoglobins
    122       2 Hemoprotein Oxidation-Reduction
    123       2 Free Energy Dependence of the Rate of Long-Range Electron Transfer in Proteins.
                Reorganization Energy in Ruthenium-Modified Myoglobin
    169       2 Biological Applications of Raman Spectroscopy. Volumes 1 and 2.  Edited by Thomas
                Spiro (Princeton University). John Wiley and Sons:  New York. 1987. Volume 1:  xv +
                343 pp. $59.95. ISBN 0-471-81573-X. Volume 2:  xi + 367 pp. $59.95. ISBN
                0-471-81574-8
    181       2 Hemoglobin-Gold Binding
    184       2 RR Spectra of O2 Adducts of Co Picket Fence Porphyrins
    246       2 UVRR Excitation Profiles of Tyrosine
    255       2 Characterization of Two Dissimilatory Sulfite Reductases
    274       2 A Deoxymyoglobin Model
    534       2 Heme d Prosthetic Group
    535       2 Multicomponent Redox Catalysts
    604       2 [Tetrakis(dichlorophenyl)porphinato]iron-Oxene Adduct
    607       2 Subsite-Differentiated Analogues of [4Fe-4S]2+ Clusters
    704       2 Access to Metal-Free Isobacteriochlorins
    752       2 NMR of High-Spin Iron(III) Porphyrins and Chlorins
    755       2 NMR Characterization of H-Bonding Networks in HRP
    850       2 Free Energy Effects on Biological Electron Transfer
    893       2 MCD of Horseradish Peroxidase Compound I
    954       2 Advances in Inorganic Biochemistry. Volume 7:  Heme Proteins.  ##aut##By G. L.
                Eichhorn (NIH Gerontology Research Center) ##aut##and L. G. Marzilli (Emory
                University). Elsevier Science Publishing Company:  New York, NY. 1987. xiv + 271 pp.
                $75.00. ISBN 0-444-00826-8
    976       2 Electronic Structure of Plastocyanin
    979       2 Low-Spin Cyanide Adduct of Transferrin
   1011       2 Mo- and W-Substituted Hemoproteins
      9       1 Bilirubin Hydrogen Bonding in Solutions
     40       1 Photoreactivation of Serine Proteinases
     44       1 Asymmetric Synthesis with Chiral b-Lactams
    101       1 Citreoviral and Citreoviridin Syntheses
    102       1 b-Carboxyaspartic Acid Synthesis
    128       1 On the Red Shift of the Bacteriochlorophyll-b Dimer Spectra
    142       1 An Unusually Stable MnIIMnIII Complex with Novel EPR Spectra:  Synthesis, Structure,
                Magnetism, and EPR Analysis###fnt#laid3cfl##
    146       1 Synthesis of the Bicyclic Core of the Esperamicin/Calichemicin Class of Antitumor
                Agents
    174       1 Thermophilic Semisynthetic Flavoenzyme
Which article number to view?
```

Sample query and title display

[redacted] Imidazole H-Bonding

Hydrogen Bonding to the Proximal Imidazole in [redacted] Model Compounds: Effects upon Oxygen Binding and Peroxidase Activity
J. Amer. Chem. Soc. vol. 110 no. 001, JA8711268

Traylor, T. G.
Popovitz-Biro, R.

Abstract

The kinetics and equilibria for binding carbon monoxide or dioxygen to the previously described adamantane-[redacted]-[6.6]cyclophane were changed very little by substituting the internally hydrogen-bonded base, 4-(2-N-piperidylethyl)imidazole for 1,5-dicyclohexylimidazole.
By contrast, the rate of reaction of protohemin dimethyl ester chloride with tert-butyl hydroperoxide was accelerated by substitution of the internally hydrogen-bonded base for N-methylimidazole.
We conclude that hydrogen bonding of the proximal imidazole increases peroxidase activity of iron(III) porphyrins but does not greatly affect oxygen affinity of iron(II) porphyrins.

Text
The affinities of five-coordinated iron(II) porphyrins for dioxygen have been shown to depend upon steric effects operating upon either the proximal (base-binding)[1-8] side or distal (O_2-binding)[8-6] side and to be increased by increases in electron donation to the porphyrin[7] or to the proximal base,[8,9] by increases in the polarity of the medium[10,11] or local polar effects,[10,12,13] and by hydrogen bonding to bound dioxygen.[12b,13,14]

Text with scroll bar, highlighting

(a) Collman, J. P.; Brauman, J.; Collins, T. J.; Iverson, B.; Sessler, J. L. J. Am. Chem. Soc. 1981, 103, 2450.
(b) Collman, J. P. Acc. Chem. Res. 1977, 10, 265.

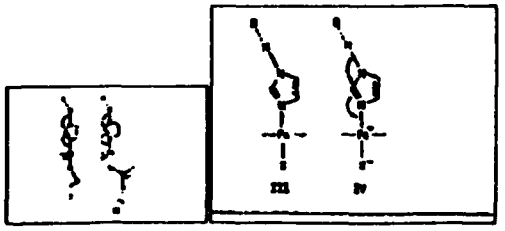
7.
(a) Traylor, T. G.; White, D. K.; Campbell, D. H.; Berzinis, A. P. J. Am. Chem. Soc. 1981, 103, 4932.

(b) Chang, C. K.; Traylor, T. G. Ibid. 1973, 95, 8475.

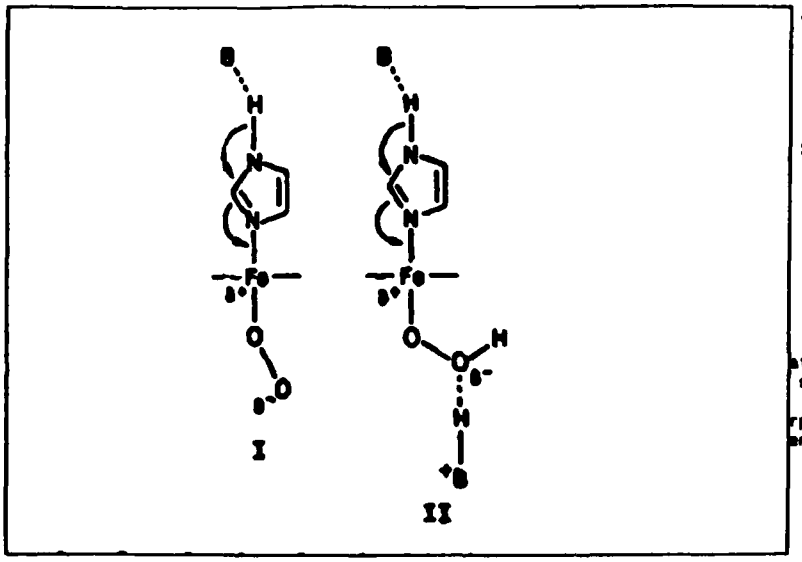
14.
Walker, F. A.; Bowen, J. J. Am. Chem. Soc. 1985, 107, 7632.

These observations, along with [redacted] crystal structures,[15,16] suggest that hydrogen bonding to the N-H of the proximal base imidazole, by releasing electron density to the [redacted] iron, should increase dioxygen affinity as seen in I. ##gid#JA1285a##
Similarly, it has been suggested that such hydrogen bonding in the iron(III) porphyrin complexes should increase the rate of heterolytic cleavage of the O-O bond (II) in proceeding to the high-valent iron species in peroxidases[17-19] and in model metalloporphyrin complexes.

17.
Paisach, J. Ann. N.Y. Acad. Sci. 1975, 244, 187.



Text, with graphics shown as icons



Am. Chem.

Soc.

at hydrogen bonding to the N-H
should increase dioxygen affinity as

rphyrin complexes should increase the
ent iron species in

Graphic expanded after mouse click

Traitement du langage naturel
Présentation d'applications pratiques

par
N. Benamou et L. Canivez
SELISA
17 Avenue du Parc
91380 Chilly-Mazarin
France

Résumé

Les techniques d'analyse du langage naturel permettent actuellement un traitement avancé de textes intégraux. En informatique documentaire, elles trouvent ainsi de nouvelles applications pour les tâches d'indexation, d'interrogation et de recherche, et enfin de traitement bibliométrique. Nous présentons dans cet article des applications réelles dans un environnement micro-informatique sur les thèmes de l'aide à l'indexation et de la bibliométrie.
D'une part, durant l'étape d'indexation d'un document dans une base documentaire, une analyse rapide d'un résumé ou d'une introduction permet d'associer au texte un ensemble de descripteurs appartenant à un thésaurus. Une analyse un peu plus fine permet l'extraction de descripteurs libres mais caractéristiques du document du point de vue linguistique.
D'autre part, tout traitement bibliométrique repose sur une analyse statistique d'un volume important de fiches documentaires. Un des apports des techniques d'analyse du langage est la mise en évidence de descripteurs libres (non définis a priori) statistiquement et linguistiquement intéressants. Ces descripteurs complètent les mots clé des fiches documentaires et offrent donc une meilleure base aux traitements bibliométriques usuels.

Introduction

La chaîne d'informatique documentaire comporte plusieurs tâches dont les plus importantes sont l'indexation, la recherche et l'interrogation. Des traitements bibliométriques et infométriques sont apparus plus récemment afin d'assurer un suivi, et non plus une exploitation, des bases documentaires.
De par la montée en puissance des micro-processeurs et de par les progrès théoriques des années 70-80 en informatique linguistique, les techniques d'analyse du langage naturel commencent à apporter des aides intéressantes et adaptées à l'informatique documentaire.
Depuis 1987, l'approche suivie à SELISA consiste à développer des outils d'aide pour les tâches de bibliométrie et d'indexation. Les applications ainsi conçues reposent sur un module de traitement du langage français, module opérationnel depuis Septembre 1989 sur micro-ordinateurs compatibles PC.

Le présent article contient tout d'abord un descriptif de l'approche linguistique retenue.
L'adaptation pratique de cette approche aux thèmes de la bibliométrie et de l'aide à l'indexation est ensuite détaillée. Des limitations sont nécessaires afin d'assurer d'une part une certaine généralité à l'outil, et d'autre part des temps de traitement raisonnables.
Enfin sont présentées deux applications réalisées dans un environnement micro-informatique sur ces deux thèmes. Ces applications, limitées au traitement du Français, sont actuellement en phase d'intégration et de test pour la base documentaire TELEDOC, base rassemblant des publications techniques dans le domaine des télécommunications.

1. L'approche linguistique retenue

Le traitement du langage nécessite d'établir une distinction entre divers niveaux de complexité de la langue : mot, phrase, discours.

Les traitements du mot consistent à en reconnaître et à en contrôler toutes les formes dérivées (féminin, pluriel, conjugaison, ...). Ceci constitue actuellement une tâche théoriquement maîtrisée qui ne nécessite que la définition de lexiques. A titre d'exemple, un lexique relativement complet du français comprend environ 70.000 mots (hors formes dérivées).

Les traitements de la phrase sont beaucoup plus complexes et se partagent entre analyses syntaxique et sémantique. Bien qu'aucune de ces composantes n'ait pour l'instant reçu de solution complète, plusieurs approches théoriques apportent des résultats exploitables.
D'une part, les approches purement syntaxiques, principalement issues des travaux de N.Chomsky (1), offrent l'avantage de n'utiliser que des

informations grammaticales. Par contre, elles ne permettent pas une gestion efficace de la synonymie et des significations voisines (Ex: "vouloir manger" est équivalent à "avoir faim"). D'autre part, les approches sémantiques visent à affiner l'analyse syntaxique d'une phrase grâce au sens des mots qui la composent. Ceci doit permettre de lever des ambiguïtés et d'associer des expressions de significations voisines. Par contre, ces approches sémantiques nécessitent d'associer à chaque terme d'un lexique une, ou plusieurs, définition. La complexité et le volume de cette tâche sont tels qu'il est illusoire d'espérer constituer un lexique sémantique important et exploitable.
Pour plus d'information sur les diverses approches de l'analyse de la phrase, le lecteur peut se référer à (2).

L'analyse du discours est encore d'un niveau de complexité tel que la plupart des travaux sur ce thème sont menés par des laboratoires de recherche. L'état actuel des recherches est décrit dans (3).

L'approche retenue à SELISA a été de développer un analyseur de phrases combinant syntaxe et sémantique. Ce programme, baptisé LangNat, exploite des informations sémantiques, mais prend également en compte des termes sans définition. Il dispose également d'un lexique grammatical de 77.000 mots (hors formes dérivées) du français général.

2. Adaptation à l'informatique documentaire

En informatique documentaire, nous nous intéressons au problème de l'identification de descripteurs à partir de texte intégral, ce qui concerne à la fois les thèmes d'indexation et de bibliométrie. Puisque, dans les deux cas, les volumes de textes à traiter sont relativement importants, il est indispensable d'effectuer principalement des traitements simples et rapides, le recours aux traitements complexes et longs devant se produire de façon limitée.

Dans cette optique, nous avons choisi un programme à deux niveaux

a) Des **traitements morpho-lexicaux** des mots, permettent d'associer très rapidement un ensemble de descripteurs à un document ou à un ensemble de fiches documentaires. De plus, certains de ces traitements sont totalement indépendants de la base documentaire considérée.

b) Un **traitement syntaxico-sémantique** permet d'affiner les résultats, mais crée quelques contraintes : temps de traitement environ 10 fois plus élevé, nécessité de connaître sémantiquement une partie importante du vocabulaire. Ceci signifie que l'adaptation de cet outil à une nouvelle base documentaire peut être long.

Afin d'obtenir des temps de réponse acceptables (en environnement micro-informatique), la coopération de ces deux modules vise à limiter le traitement syntaxico-sémantique à l'analyse de descripteurs potentiels.

Dans ce contexte, le premier module traite l'ensemble d'un document à indexer ou de résumés documentaires téléchargés et propose une liste de descripteurs morpho-lexicaux. Il s'agit de groupes de mots, formes de base et non formes dérivées, tels "réseau .. connexion" ou "transmission large bande". Cette approche morpho-lexicale offre bien entendu une certaine finesse dans l'identification du vocabulaire et donc des expressions clés. Par contre, elle ne permet pas la mise en évidence de synonymie entre mots ou expressions. Elle ne peut mettre automatiquement en rapport des textes comme :

-Commande du réseau de signalisation
-Interface d'adaptation de signalisation

Pour établir ce type de comparaison, le système morpho-lexical devrait autoriser la prédéfinition d'un ensemble d'expressions bien trop important pour être exploitable aisément.

Pour chacun de ces descripteurs potentiels, le second module effectue ensuite un traitement syntaxico-sémantique de l'ensemble des phrases le contenant. L'objectif de ce module est de construire une représentation de la syntaxe et de la sémantique - en d'autres termes: de la structure et du sens - permettant un codage du texte, codage relativement indépendant du vocabulaire précis utilisé.
Ainsi les deux expressions
-Commande du réseau de signalisation
-Interface d'adaptation de signalisation
ont une partie commune :
-Moyen de modification de signalisation
qui peut être automatiquement générée. Cependant, la détermination de cette partie commune étant une opération relativement complexe et longue, il convient de la limiter à des expressions statistiquement importantes.

En résumé, cette approche à deux niveaux permet de conserver des temps de traitement raisonnables tout en effectuant sur des points particulièrement ciblés une analyse fine.

3. Application à la Bibliométrie

La bibliométrie a pour objet l'analyse de fiches documentaires déjà stockées dans une base documentaire afin d'en tirer des tendances et des informations statistiques. Dans le cadre d'un projet mené par un organisme public de documentation, une analyse bibliométrique est menée selon le protocole suivant :

- Téléchargement sur un micro-ordinateur d'un ensemble limité de fiches documentaires (environ 100) au moyen d'une interrogation classique.
- Extraction de descripteurs par analyse des champs textuels de ces fiches afin de compléter le champ Mots-Clé.
- Analyse statistique de ces fiches en prenant pour base les mots clé et les descripteurs additionnels.

Le traitement du langage naturel permet d'affiner l'étape d'extraction de descripteurs. Sur ce thème, des tests d'intégration sont actuellement menés sur la base TELEDOC.

Certaines caractéristiques des fiches documentaires téléchargées orientent fortement les traitements linguistiques. Tout d'abord, les champs textuels contenus dans les fiches étant rédigés par des analystes, et non pas par les auteurs, le style de rédaction est fréquemment télégraphique, marqué par des phrases courtes, énumératives ou sans verbe. Ensuite, les fiches étant téléchargées à l'issue d'une session de consultation, elles traitent d'un domaine assez restreint.
Enfin, les fiches étant le résultat de l'indexation de documents, les termes contenus dans le champ Résumé et appartenant à un thésaurus sont généralement inclus dans le champ Mots-clé.

De ces caractéristiques, l'on peut tirer deux règles de conduite :

- Il est inutile d'utiliser un thésaurus du domaine technique pour rechercher des mots-clé.
- Combiner une analyse morpho-lexicale et des traitements statistiques doit conduire à des résultats pertinents. L'appel à l'analyseur syntaxico-sémantique peut donc être très limité.

Ainsi sur un ensemble de 100 résumés documentaires extraits de la base de données TELEDOC sur le thème de : la *commutation de données et les transmissions par satellite*, le programme informatique propose un ensemble d'une centaine de descripteurs commençant par :

fibre optique	51
réseau local	47
système télécommunication	14
modulation déplacement	13
transmission numérique	12
bande base	11
service offrir	10
signal bruit	10
réseau numérique	9
réseau connexion	9
système transmission	9
liaison montante	9
modulation MDP	9
réseau commutation	8
liaison satellite	7
transmission donnée	7
numérique satellite	7
réseau commuté	7
...	

Diverses stratégies sont en test à SELISA afin de choisir les descripteurs à affiner par une analyse systématique de toutes les phrases les contenant. Les trois plus prometteuses sont les suivantes :

- Les descripteurs apparaissant plus d'un certain nombre de fois (entre 5 et 10) peuvent être analysés.
- Les descripteurs appartenant déjà au thésaurus ne nécessitent pas d'analyse complémentaire.
- Les descripteurs présentant une structure grammaticale simple : NOM + ADJECTIF ne nécessitent pas d'analyse complémentaire.

Compte tenu de ces stratégies, l'analyse de l'ensemble des phrases contenant les descripteurs sélectionnés conduit à un nouveau jeu de descripteurs dont les mots constitutifs peuvent être différents. Ces nouveaux descripteurs reflètent une structure syntaxico-sémantique et non plus une proximité de mots dans une phrase. Dans l'exemple précédent, ceci se traduit par la création des expressions :

transmission numérique par satellite
réseau commuté

à la place de

numérique satellite
réseau commutation
réseau commuté

En conclusion, le module de traitement syntaxico-sémantique vient modifier, épurer et enrichir la liste des descripteurs trouvés en première analyse. Ces descripteurs sont alors affectés aux diverses fiches documentaires. Pour ce faire, un fichier, initialement une copie du fichier téléchargé, est enrichi d'un champ DESCRIPTEUR, ce champ étant ajouté à la fin de chaque fiche.

Cette approche permet d'insérer simplement, et de manière optionelle, ce module de recherche de descripteurs dans une chaîne de traitement bibliométrique. Les analyses statistiques peuvent alors être indifféremment appliquées soit au fichier téléchargé, soit au fichier complété.

4. Aide à l'indexation

L'indexation d'un texte dans une base documentaire passe par la définition de mots clé représentatifs du texte. Actuellement, cette tâche est réalisée par des analystes professionnels, en grande partie de manière manuelle. Dans la mesure où nul ne possède une culture universelle, de nombreux documents techniques restent obscurs aux analystes et sont donc mal indexés. De plus, pour les bases associées à des thésaurus, certains descripteurs appartenant au thésaurus sont omis lors de l'indexation (un thésaurus pouvant atteindre 100.000 termes).
En conséquence, un outil basé sur l'analyse du texte intégral permettrait de simplifier grandement la tâche des analystes.

Contrairement aux fiches documentaires, les textes concernés par l'aide à l'indexation (articles techniques, ...) sont structurés et rédigés dans un style relativement élaboré. Ensuite, ces textes sont susceptibles de toucher tout thème couvert par la base documentaire dans laquelle un analyste cherche à les intégrer.
Ces deux caractéristiques, opposées à celles rencontrées en bibliométrie, impose une nouvelle ligne de conduite pour un logiciel d'aide à l'indexation :

- La première tâche à accomplir est la recherche, au sein du document, de descripteurs appartenant à un thésaurus.
- A l'issue de cette étape, une analyse morpho-lexicale et des traitements statistiques permettent la mise en évidence de descripteurs potentiels qui seront validés par une analyse syntaxico-sémantique. Si la recherche de mots-clé dans le thésaurus est satisfaisante, l'analyste peut ne pas effectuer cette seconde étape.
- Les descripteurs libres obtenus à l'étape précédente peuvent être introduits dans le thésaurus, la décision étant du ressort de l'analyste.

L'originalité, par rapport à l'application bibliométrique présentée en 3., est donc l'utilisation d'un thésaurus "*dynamique*" comme premier niveau d'analyse. L'objectif recherché est que ce thésaurus s'enrichisse suffisamment pour rendre inutile, ou rare, la seconde étape d'analyse. Le thésaurus ainsi constitué doit contenir deux types d'informations :

- L'organisation du descripteur comme groupe de mots, e.g. "transmission numérique satellite".
- La structure syntaxico-sémantique associée à ce descripteur. Cette structure permettra de gérer les synonymies de mots ou de groupes de mots.

Dans cette optique, SELISA réalise actuellement une extension du logiciel utilisé en bibliométrie. En l'état actuel des développements, l'étape de recherche des descripteurs appartenant au thésaurus est achevée et interfacée avec le module bibliométrique. Par contre l'étape d'enrichissement du thésaurus n'est pas encore implémentée (fin prévisionnelle en Décembre 90). La version actuelle permet toutefois de se rendre compte de la complémentarité de la recherche de descripteurs par comparaison avec un thésaurus d'une part, et par analyse linguistico-statistique d'autre part.

Quel que soit l'intérêt d'un logiciel d'aide à l'indexation, il convient de garder à l'esprit que son domaine de validité est limité à des documents stockés informatiquement sous forme "caractère" et non pas "image". Ceci impose de :

- Soit traiter exclusivement des documents issus de traitements de textes ou d'outils de rédaction informatisée.
- Soit mettre en oeuvre un système de saisie des textes à indexer. Les deux solutions possibles sont la saisie manuelle ou un scanner associé à un système de reconnaissance des caractères. Quelle que soit la solution retenue, il s'agit là d'une opération relativement lourde et complexe à gérer.

Dans ces conditions, l'aide à l'indexation apparaît comme un service en devenir, dont l'expansion sera essentiellement liée à celles de la micro-informatique et des messageries électroniques.

Conclusion

Après plusieurs décennies pendant lesquelles le traitement du langage naturel a été considéré exclusivement comme un sujet de recherche, des débouchés pratiques apparaissent, en particulier en informatique documentaire. Nous avons présenté ici deux applications, une en phase d'intégration et de test pour des traitements bibliométriques, l'autre en phase de développement pour l'aide à l'indexation. Il faut également noter qu'une application tout aussi importante en informatique documentaire serait l'interrogation en langue naturelle des serveurs ou des bases. Par extension, ce thème concerne également l'accès aux bases de données non-documentaires.
Toutefois, une des limites essentielles de ces applications est la nécessité d'utiliser une, et non plusieurs, langue. Ainsi, les applications pratiques

présentées précédemment ont été conçues pour le Français et ne permettent pas d'utiliser des serveurs documentaires anglo-américains.

Références

(1) Chomsky N. "Aspects of the theory of syntax", MIT Press,1965, Cambridge, Massachusetts

(2) Desclès J.P. "Tendances de la linguistique contemporaine : Différentes notions de la grammaire universelle", Actes du séminaire Recherche et Développement dans les Industries de la Langue, INRIA, 1987, Rocquencourt, France, pp55-114

(3) Sabah G. "L'intelligence artificielle et le langage", volume 2, Hermes,1989, Paris, France, pp187-316

BEYOND MACHINE TRANSLATION

by L. ROLLING
Computerized language processing unit
Commission of the European Communities
Jean Monnet Building B4/29 - L-2920 Luxembourg

The author describes the essential aspects of computer translation : design, funding, development, experimental use, evaluation and implementation. He emphasizes the technical, economic and psychological obstacles to be overcome for efficient use of the new instrument.

Current applications will be supplemented by other, new applications in the MT market, which may well be centered on commercial communication and data base access.

While the technical feasibility and the economic viability of MT depend on system design and implementation infrastructure, the quality of MT output ultimately depends on the representativity of available grammars and dictionaries vis-à-vis the real-life use of human language.

But MT is not the only application that requires a complete mastery of a number of languages by computer. Publishing, data base maintenance and interrogation, and many other natural-language processing activities require all-encompassing text corpora, multi-purpose lexica, term banks and other linguistic and computational resources.

The EC Commission has a major rôle to play in the coordination of European efforts to standardize, develop and manage these resources, to be used also for speech-technology and knowledge-based applications of the future. EC action includes a number of ESPRIT and IMPACT projects, but the coordinating and teaching tasks, now under the Commission's Multilingual Action Plan, may well be handled by a new agency to be created for this purpose.

1. ESSENTIAL ASPECTS OF MT

Automatic translation was invented one generation ago and it has attained today a level of quality that allows us to say that it is here to stay, in spite of the declarations of pessimists who say that it is not automatic because the results require some monitoring on behalf of the end user.

If you ask a user why he uses MT, most of the time he will not mention the translation quality or even the reduced cost, but he will refer to the increase in speed. What it took a human translator two weeks to provide, he can now obtain within ten minutes after pushing a button (or ten seconds if he uses Minitel for the translation of a short paragraph).

However, quality, low cost and speed are not enough. What the end user needs is the push-button device, i.e. the compatible equipment that allows him to input the source text, be it on an ASCII diskette or on paper, in commercial print format or in low quality typewriter script, that allows him to steer the operation towards the desired target language and format, using the terminology of the relevant subject field, and through the mainframe and back to his printer.

The need for these indispensable tools was often ignored by the supplier, and the user was told that his secretary had only to type in the source text, that he had only to sit at the screen and answer the system's disambiguation questions, and that he had only to provide a printer capable of producing acceptable output.

Today there are quite a number of MT suppliers, and they have become aware of the need not only to supply the user with an efficient software product, but to help him to install compatible hardware, to introduce his specific terminology into the dictionaries, to make his work station as user-friendly as possible and to make the best use of his feedback for system improvement.

The suppliers have realised that they can only sell or license their systems if the first users are satisfied and let it be known. If they want to have a large number of satisfied clients, they have to take account of a variety of requirements, concerning source and target languages, specific terminologies, and work stations adapted to the users' computer infrastructure and global environment.

All this requires considerable investments, and the suppliers should see to it that the development, marketing and operating costs are reduced as far as possible.

One way to achieve this is through stratified or multi-level dictionaries. Dictionaries can be subdivided into (dominant) personal or local dictionaries and (default) universal dictionaries; frequently occurring words can be given precedence over occasional terminology, etc....

It is also possible today to automatically produce multilingual dictionaries starting from equivalent texts in several languages such as the EC Official Journal, the proceedings of the European Parliament and Canadian, Belgian and Swiss legislation.

(see Table 1)

2. EXISTING SYSTEMS

The first MT systems were developed in the sixties on mainframe computers that were less powerful than today's personal computers. Many have disappeared since, and the only success stories concern SYSTRAN, LOGOS, METAL and SPANAM. Systran users include Xerox, the US Air Force, NATO and the EC Commission.

LOGOS is used by a number of Canadian organizations and SPANAM is owned by PAHO, the Pan American Health Organization. METAL, designed by Texas University, is now marketed by Siemens.

Quite a number of systems were developed later for the emerging generation of desktop computers. They are interactive rather than autonomous, i.e. the user has to be present to help the system solve ambiguities. Other systems such as ALPS and INK are just translators' aids.

TITUS and TAUM are specific in that they require restricted syntax and terminology.

Much remains to be said about Japanese research efforts, which have already led to almost a dozen operational systems for translating from and into Japanese, and about European research projects, such as the European Community's EUROTRA project and two Dutch initiatives launched by Philips and BSO.

The most interesting initiative, however, emanates from the USA, where Carnegie-Mellon University is plunging into the unexplored depths of semantic (as opposed to morpho-syntactic) analysis.

(see Table 2)

3. FUTURE OUTLOOK

Looking at today's literature on the subject, one can spot a number of fascinating ideas. Artificial intelligence, applied to MT, will solve the semantic problem of ambiguity resolution, as soon as the necessary knowledge bases come into existence. Neural networks will allow for parallel and interlinked computing, so that the most cumbersome analysis systems will be speeded up by several orders of magnitude. And speech technologies linked to MT will develop a "telephone interpreter" by the year 2000.

How should we take such pronouncements? Well, that major progress of these types never occurs suddenly, "en bloc".

Many experts will spend many years developing viable knowledge bases, and unfortunately, machine translation might well not be their major priority. The same is true for neural networks and parallel computing : once these technologies have been developed for other, more important projects, MT specialists might start to learn how to apply them to their own environment. Voice technologies have a closer link to MT. Speech generation is technically viable today, but speech recognition and understanding is at least one decade away. Fortunately, it will be possible to design interactive systems, allowing a speaker to monitor his utterances on a screen before launching them into the MT system.

Both voice analysis and synthesis require very large phonological dictionaries and speech data banks for the development of multi-speaker, speed- and noise-independent systems.

In the meantime, we have to do a lot of dirty work that does not require intelligent research capabilities, but consistent efforts by competent linguists and programmers.

Many systems that had been developed under sophisticated, academic programming languages like LISP or PROLOG, will have to be converted to more efficient devices; MS-DOS systems will be transferred to UNIX, ALGOL systems to ADA, etc.....

Electronic dictionaries developed for publication and/or NLP applications can be made re-usable for MT systems. Mono- and multilingual text corpuses can be exploited to produce terminology and equivalent sublanguage patterns, making development and use of MT cheaper and cheaper, while human translators, in insufficient numbers to perform the tasks that are definitely not for MT (literature, publicity, speeches), will increase their prices, making MT even more competitive.

Translations will be stored in large corpuses, and retrieval of already translated text chunks will become competitive as repetitive translation of everyday texts increases over the years.

Improved MT work stations will be developed to give the operator access to previous translations, to lexical and terminological resources not yet available in the MT dictionaries, and to allow him to produce a variety of products and services in various languages, formats and scripts, so as to satisfy an increasing array of impatient customers.

(see Table 3)

Table 1

ESSENTIAL ASPECTS OF M.T.

1. DESIGN	- Direct (bilingual) - Transfer - Interlingua - Pivot language
2. ECONOMICS	- Market (open, hidden) - Funding (public, private) - Viability threshold - Maintenance cost
3. EVALUATION	- Quality (revision rate) - Speed (CPU, turnaround) - Cost (raw, post-edited) - User-friendliness
4. IMPLEMENTATION	- Input (OCR, Spell check) - Text typology (correctness) - Computing / Networking - Post-editing (format, replace)
5. HUMAN ASPECTS	- Development - Training (authors, users) - Acceptance / Promotion - Management (feedback)
6. FUTURE PLANNING	- Resources (corpora, lexica) - Automation (extraction, retrieval) - Speech input/output - YOU NEVER KNOW

Table 2

EXISTING SYSTEMS

Mainframe systems :	LOGOS SYSTRAN PAHO METAL
Desktop systems :	SMART WEIDNER-BRAVICE GLOBALINK D'AGOSTINI TOVNA
Task-specific systems :	TITUS TAUM-Meteo
Translators' aids :	ALPS ERICSSON INK
Japanese systems :	ATLAS (Fujitsu) HICATS (Hitachi) MU (Kyoto-JICST) TRANSAC (Toshiba) MELTRAN (Mitsubishi) etc.....
Research projects :	EUROTRA (EC) BSO/DLT ROSETTA CARNEGIE-MELLON + Japan + IBM ...

Table 3

Language Engineering Programme

Preparatory period : 1989-91

Programme period : 1992-94

Current activities

- Inventory of current products and services, research activities and projects.
- State-of-the-art studies for various subsectors of language industry.
- Economic impact studies for NLP, MT and speech technologies.
- Definition of priorities for coordination : common formats and standards.
- Development and coordinated exploitation of
  - . representative text corpora,
  - . lexical resources,
  - . terminological resources,
  - . software and hardware products,
  - . phonological dictionaries and speech data banks.
- Creation of a European Institute for Language Engineering.

THE ROLE OF INTELLIGENT ONLINE INTERFACES TO BRIDGE THE COMMUNICATION GAP

by
A VICKERY
TOME ASSOCIATES LIMITED
IMO House
222 Northfield Avenue
LONDON, W13 9SJ

ABSTRACT

People are curious from nature, want to know more than they do, have an insatiable gluttony for information, whether in their pursuit of private aims or in their professional career. The best carrier of information is person to person communication. A postgraduate student will seek information from his tutor or professor; apprentice - from his factory supervisor; a working engineer from his colleagues. A small proportion of people try to get information by reading and even a smaller number will try to access online databases.

And yet, very great deal of technical and professional communication today, is mediated through documents, or even more indirectly, through computer systems. Communication gaps between man and system are just as real and important as any other, and it is this kind of gap that will be considered here.

The paper will first enumerate the difficulties which arise in accessing computer-based information:

1. the language barriers such as databases in different national languages, concepts having different meanings in different databases (or parts of the same database), variations in command or query languages.

2. the intellectual difficulties, i.e. the gap in knowledge which exists in the searcher's mind during the stage of search formulation, the misunderstandings which can arise during the human/human communication (if the search is done by an intermediary), and errors arising from human/computer communication during the search process.

3. the technical barriers in achieving a satisfactory search result, i.e. in communication with various hosts and many databases, in using telecommunication links, in different techniques in interrogating files, in different indexing methods, in variations in structure of vocabularies, in classification.

The conclusions of the paper will summarise the current achievements in overcoming the barriers to information for online databases and the problems which still need solutions.

1. INTRODUCTION

The number of databases and electronic information services publicly available in Europe, USA and elsewhere is now large (4300 according to Cuadra) and continues to grow. But actual usage of online databases remains low in comparison to the number of professionals of all kind which could benefit from information that is accessible. Below we shall consider steps to be taken by a user to carry out a successful search. In all these steps one can see barriers which may prevent the user from using online searching.

2. THE ONLINE SEARCH PROCESS

The databases to which electronic information services give access are of various kinds. A useful categorisation has been provided by Staud (1988), who recognises the following types:

1. The factual databases
   Statistical, with processing facilities
   Quasi-statistical, tables but no statistical processing
   Textual facts, including referral, directories
   formalisms (eg. chemical structures), models
2. Textual
   Full-text
   Bibliographic
3. Integrated, eg. combining textual facts, tables and a bibliographic reference

Whatever the type of database, access in nearly all cases to its data is via an index that consists of words, phrases, names, codes, class numbers, numerical identifiers, citations or other elements in the records. Each database consists of records and all records have fields.

Search is primarily carried out on the data contents of the record within its various fields. Search terms can be single words, phrases or codes. These can be combined into Boolean structure by the use of the operators AND, OR, NOT. A simple use of the structure of textual discourse is represented by proximity operators; words can be sought that are adjacent in the text; or near each other; or in the same field etc. Some use is also made of the semantics of individual words - words with common elements of meaning can be conflicted by truncation (right-hand, left-hand, or internal). Search may be

limited to records in a certain language or a certain date.

To undertake a search, the user must take decisions. Below are listed the most important ones:

1. The user must have access to a terminal (or microcomputer acting as such)
2. The microcomputer must be linked via telecommunication to a variety of mainframe (or minis) online hosts, on which are mounted databases. For this purpose the user must have a signed contract with the national telecommunication agency and the hosts s/he will use.
3. The user must be aware of question s/he wants to put to the system. It is not such an easy task to formulate a query in the area in which a knowledge gap exists in the user's mind. The answer to the query must fill in this gap.
4. The user must know the databases s/he wants to interrogate and the hosts on which they are available.
5. Query is formulated in the vocabulary suitable for the selected databases (this may involve use of a thesaurus).
6. Query is expressed as a search statement in the format required by the selected hosts, using Boolean and other search operators (proximity, truncation, field restriction, limits).
7. Selected host is dialled up via telecommunications and logged on.
8. Selected database file is entered.
9. Search statement is transmitted to host, using appropriate command language to instruct the mainframe computer what to do. (Command language for each host is different)
10. Search output is presented to user in selected format.
11. If search output is not acceptable to the user, search statement is amended and processed again as in 9 and 10.
12. Switching takes place between databases, if required.
13. Search output is delivered to user (printed or downloaded).
14. Further hosts may be accessed.
15. Documents may be ordered online.

From the above list of functions performed during an online search, it becomes evident how many skills must be acquired before a successful result can be achieved. An information intermediary can help the user in many aspects of the problem. But not each institute can afford to employ enough of these skilled practitioners to satisfy its workers and there are many medium and small firms which have not even one single intermediary.

The concept of an 'intelligent interface' tries at least to some extent to help the end-users. An 'intelligent interface' is a software package that is interposed between the searcher and the database system, and that can provide the user with some (ideally all) of the help that an intermediary gives. What do we mean by the word 'intelligent'? It is used to mean any software package that replaces any action normally undertaken by an information intermediary in online database search. These actions can range from the purely clerical , such as automatic dial-up of telephone numbers, to the fully intellectual, such as selecting the best way of modifying an unsuccessful search query or questioning the user to clarify an input query.

Let us see how an intelligent interface can help the end-user in the previously described functions. A number of intelligent interfaces have been developed over the last few years, most of them only to the prototype stage (B.C.Vickery, 1989). I shall be mentioning one or two of these, but most of my comments will be illustrated with reference to a product with which I am associated, the TOME.SEARCHER. Its procedures are being improved and incorporated in MITI, an European Community project in the IMPACT programme.

3. HELPING WITH USER QUERIES

The user inputs a query to the system. An 'intelligent' system should assist the user in 3 aspects:

(i). The best mode of communication between the end-user and the machine, in either direction, would be a natural language interface. Can it be easily achieved?

(ii). Is the search formulated in a satisfactory way so that the system can supply the user with an efficient answer?

(iii). User models can enable the system to take into account the differing needs, skills, knowledge and expertise of groups of users or even individuals. How far did we go in development of good user models in the human-computer interfaces?

(i) Natural language

Users find it much easier to express their needs in their own language than to learn a controlled language such as many indexers use. When a natural language interface is envisaged for an information system, this interface must be able to translate the natural language statements into appropriate terminology used by the system.

Speaking about natural language interfaces, it should be realised that what is meant is just a subset of a natural language, a subset which will closely correspond to the domain covered by the information system.

The activity of processing natural language is usually divided into four phases: morphological, syntactic, semantic and pragmatic.

The morphology of a language has to do with the make-up of words, and in particular with the suffixes and prefixes that can be combined with a given 'root' word: so the root 'love' can appear as 'loves', 'loved', 'lover', 'unloved'. Many of these prefixes and suffixes are subject to regular rules and these rules can be utilised in language analysis.

The syntax of a language is concerned with the ways in which words are combined into larger units - phrases, clauses, sentences. Words play various roles in a sentence, the 'parts of speech'- nouns, verbs, adjectives, prepositions and so on.

Semantics is concerned with the underlying meaning of text. To understand a sentence fully, it is necessary to grasp not just the grammatical role of each word, but also its semantic role.

Lastly we come to pragmatics. This is concerned with the context in which a particular linguistic statement occurs.

The processing of user queries input to an information system requires morphological, syntactic, semantic and pragmatic analysis.

A system accepting free expression of a user query must process it so as eventually to create a search statement. The input string is first separated into words. Most systems for searching electronic information services remove non- significant words by scanning against a stoplist. Stopping is usually followed by stemming and the ensuing processing is carried out with the stems.

Compound terms are often identified by matching input against a dictionary, lexicon or index. TOME.SEARCHER, as well as identifying compounds that occur in its dictionary, also employs semantic rules to recognise and create other compounds appearing in the input.

In each text one can find words which have more than one meaning, eg. cell; is it a biological cell, an electric cell or a prison cell. These multimeanings must be disambiguated and some system as CIRCE do it by asking the user, some try to develop rules (eg.ERLI).

TOME.SEARCHER uses several methods including checking a multimeaning term against the context, that is the subject area in which the search is to be carried out. TOME.SEARCHER also adds synonyms to terms in the dictionary, so that they can be employed in search strategies.

So far we have spoken of language processing in one language only. But one of the more serious barriers to communication is the inability to understand a foreign language. We can break this barrier by introducing a multilingual interface to databases. The characteristics of such an interface could be as follows:

(1) screen displays can be available in all the languages covered by the system; this would include the process of database selection.
(2) the interface can accept input of user query in each of these languages and refine the query by interacting with the user in the language of input.
(3) the terms in the refined query could be translated into the language of the selected database immediately before formulation of a Boolean search statement.
(4) a final bonus would be the translation of retrieved records into the language of the user, though here we are going beyond the immediate functions of a search interface and into text translation program.

The first three features will be introduced into the MITI interface previously mentioned.

(ii) Search formulation

Everybody who ever worked with users knows how important it is to clarify the search query posed by the inquirer. In a discourse between the user and a knowledgeable, skilled intermediary there is a process of knowledge enhancement on both sides: the user specifies more tangibly the missing information and acquires at least some knowledge of databases, thesauri and vocabularies and the intermediary gains more specific knowledge in the subject area of the user. For both of them it is an excursion into each other's mind. This can be represented in a form of a Venn diagram.

Can such communication pattern be ever achieved by the man and a machine? In TOME.SEARCHER we tried to create a modest model of a discourse, using the results of semantic analysis. The problem statement is made up of a set of frames, one frame for each term. Frames consist of slots, places where a particular item of information fit within the larger context created by the frame. The slots for each term are defined by semantic categories. By supplying a place for expected information and thus creating the possibility of recognising that information is missing or incompletely specified, the slots mechanism permits reasoning based on confirmation of expectations - 'filling the slots. ' The procedure is as follows: first, the system attempts to fill as many slots as possible with information the system already possesses. Then the system checks the frames for 'sufficient completeness'. I.e. the minimum information about a particular concept required by the search strategy construction process. The rules specify which combination of filled slots is necessary and sufficient for the system to carry out a search. If it is determined that more information is required, the user is prompted to supply the missing information.

This method of eliciting information from the user worked quite well in a small system. But when similar procedures were developed for a big

database, the number of questions to the users has extrapolated considerably creating difficulties in implementing the technique. It is not known to me if a better system of eliciting information from users exists in a working 'intelligent' system.

(iii) User models

Future 'intelligent' user interfaces must be dynamic. It means that they must adjust to user requirements. Some users have never previously used online searches, they do not know how to formulate search strategies or how to use the command language. For them the search process must be automatic, a 'black box' with input and output mechanisms. Other users are experienced searchers. They need to be given the flexibility of changing a search in whatever way they want.

The user model should be able to adjust the actions of the 'intelligent' system to the individual needs of the user. There are static characteristics of the user, such as age and sex, and dynamic characteristics which will change with usage of the system, with the change of a job or even with the change of the project on which the user is currently engaged. A reasonable approach therefore would be to ask the user few questions representing his/her static characteristics which could be used permanently by the system and questions which would elucidate the ongoing changes each time the user enters the system.

Despite many attempts to implement a user model into an intelligent information system, there is little evidence so far of success.

TOME implemented user models in few of its systems. The best results have been obtained by asking the user about his location and language to be used during the interactions in the system. The location gives, for example, a display of libraries and their sources of information close to the place of living. The language, quite understandable, enables the user to understand what is going on during the search. In an intelligent tutoring system the student model acts as a historical record of what the student knows and what he needs to learn. The abilities of the student can influence the presentation of the teaching material and the testing of his achievements.

4. HOST AND DATABASE SELECTION

This is another area in which an 'intelligent' interface can assist the user.

Once an intelligent interface begins to cover heterogeneous databases with different contents and structures, it becomes necessary to keep information within the interface on how to map queries to those information sources. Similarly there is a need to keep information on the different hosts and their facilities and command languages.

Some hosts provide aids to database selection - for example both DIALOG and ESA/IRS permit the user to specify a broad subject field, and to make a trial search of the databases that the host has allocated to the subject, to discover what output each will yield on the search topic. Alternatively, databases can be selected by an interface before going online. The choice of a subject field can be guided by display of menus leading down from the broadest subjects to more specific ones; information about the databases allocated to the chosen subject can be displayed to help in final selection. Similar facilities for database selection are offered by gateways such as EASYNET.

In TOME.SEARCHER the database selection module interacts with two other modules, the user profile and source data dictionary.

The user profile besides of other data contains the host passwords which the user is entitled to use; the source data dictionary includes data on host (such as command language of the hosts, SDI facilities, printing formats) and data on each database (eg. type of database, searchable fields, language of database). The database selection module supplies descriptions of databases including subject coverage.

From the search query, the system identifies the terms and compounds by language processing. The classification numbers of the terms are then used to climb the classification of the system to converge on subject classes that have been used in database selector. Hosts and databases for which the user has no password are eliminated; the remaining databases are checked in the source data dictionary and again those that do not match search specifications (previously collected from the user) are discarded from the valid list of databases. The final list of host and databases is displayed for user approval.

This way the user is helped in choosing the right databases, being sure that none are missed out but at the end, the final choice belongs to his/her decision.

5. THE SEARCH PROCESS

The process of creating search strategy can be divided into three stages: the initial setting up of the strategy, the modification of search strategy and its translation to the host command language.

In a conventional search (without an intelligent interface), the user needs to know: the terms, their relationships with other terms, Boolean operators and the command language which can and usually does differ for each host.

In an intelligent database, the user should not need this knowledge. The terms and their relationships are determined in the dictionary of the system; the

Boolean operators are implemented mainly by the intelligence of the system, by interaction of the system supplemented by interaction of the system with the user. The created search strategy is then translated into the command language of the host (stored in the source data dictionary mentioned before). The search strategy is then released to the host database(s) and the yield of the search is presented to the user. Very often some modifications to the search are needed and these will be again introduced either by the system or by the user. Some of the possible amendment procedures are already available on some hosts - for example, ESA/IRS has a ZOOM procedure, and INFOLINE has a GET procedure.

The interface may also initiate a relevance feedback procedure: presenting sample output from the first search to the enquirer, asking him/her to evaluate each item, and using the response to amend the search statement. More weight may be given to the index terms of relevant items, and less weight to those of items judged not to be relevant.

The automatic procedures in search strategy are carried out by TOME.SEARCHER for inexperienced users. Other interfaces introduce also some help in building search strategies, in most cases with user intervention.

6. COMMUNICATION PACKAGES

There are many communication packages in use in information retrieval systems. The intelligent interface incorporates the communication package into the system. No initiative of the user is necessary, the search strategies in the appropriate command languages pass smoothly from the terminal to the host computer and back to the user terminal.

7. CONCLUSION

Intelligent interfaces can offer many facilities to make it easier to interact with online databases. I have particularly mentioned:

- helping the user to select appropriate databases
- accepting queries in the user's own language, even when it is not the language of the database
- intelligently processing these queries to formulate an optimum search
- automatically putting the search into the form required by the online system
- handling automatically all telecommunications with the host system

An interface with these facilities can go a long way to bridge the communication gap between end-users and online sources of information and overcome some of those barriers which prevent the end-user to use frequently and efficiently the online databases.

8. REFERENCES

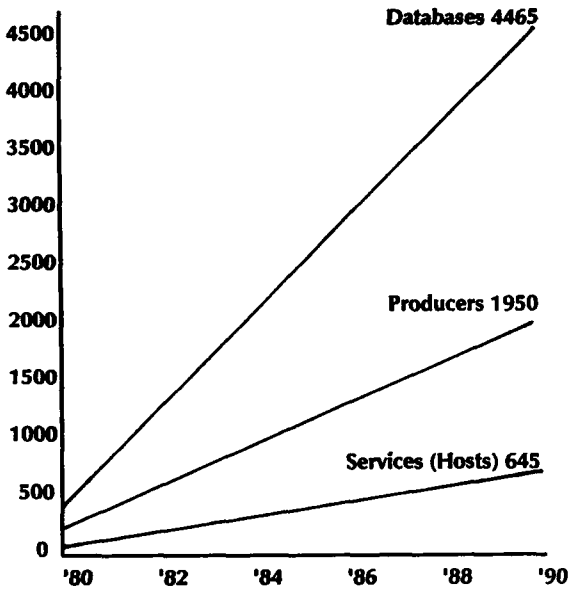
1. STAUD, J.L. (1988), The Universe of Online Databases, Journal of Information Science, 14, 141-58.

2.VICKERY, B.C. (1989), Intelligent Interfaces, State-of-the-Art Survey, CEC contract ML-60.

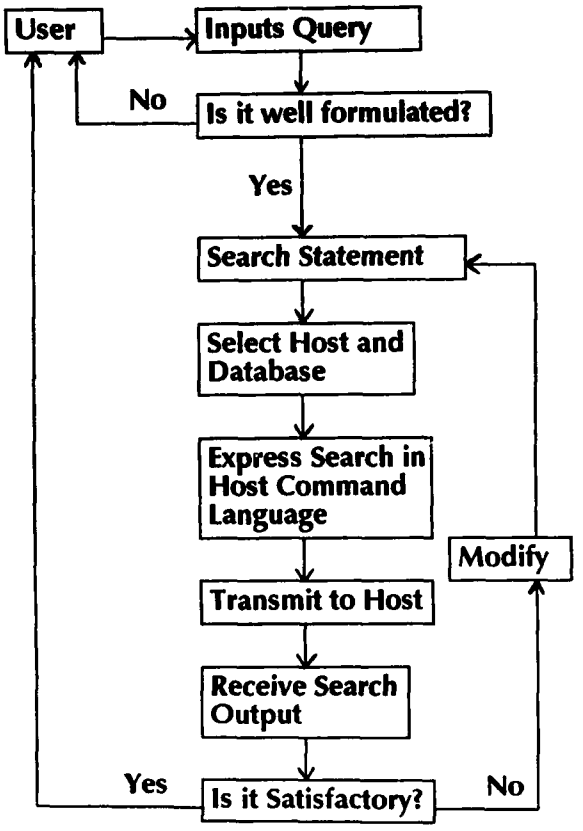
NOTE

In response to requests, the viewgraphs used during the presentation are included here (pages 8-6 to 8-9).

ONLINE DATABASE GROWTH



SEARCH ACTIONS AND DECISIONS



TYPES OF DATABASE

FACTUAL
- Statistical
- Textual Facts
- Formalisms

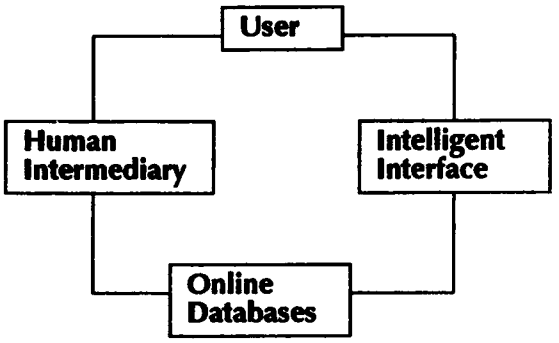
TEXTUAL
- Full Text
- Bibliographic

INTEGRATED
- eg Textual Facts
- + Tables
- + Bibliographic References

KNOWLEDGE NEEDED

1. **SUBJECT DOMAIN:**
   - Terminology
   - Thesaurus Relations
   - Classification
2. **HOST CHARACTERISTICS:**
   - Command Language
   - Search Facilities
3. **DATABASE CHARACTERISTICS:**
   - Subject Scope
   - Type of Information
   - Field Structure
   - Language
   - Output Formats
4. **TELECOMMUNICATIONS PROCEDURES**
5. **SEARCH STRATEGY:**
   - Query Formulation
   - Search Modification

SEARCH PATTERNS



LANGUAGE PROCESSING

1. Removing Non-Significant Words from Query by use of a Stoplist
2. Stemming to Remove Plurals and Other Suffixes
3. Look up in Dictionary
4. Questioning User to Clarify Words not in Dictionary
5. Creating Compound Terms
6. Resolving the Meaning of Ambiguous Words
7. Linking Query Terms to Synonyms

HELP WITH USER QUERIES

1. User can express query in natural language
2. System helps user to create a well-formulated query
3. User model guides the interaction between user and system

FROM QUERY TO SEARCH

"Papers about the handling and defects of floppy disks"

STOP:	HANDLING - DEFECTS - FLOPPY DISKS
STEM:	HANDLING - DEFECT - FLOPPY DISK
COMPOUND:	HANDLING - DEFECT - FLOPPY (w) DISK
SYNONYMS:	(HANDLING OR CARE) — (DEFECT OR FAULT) — (FLOPPY (w) DISK OR FLOPPY (w) DISC OR DISKETTE)
BOOLEAN:	(HANDLING OR CARE) OR (DEFECT? OR FAULT?) AND (FLOPPY (w) DISK? OR FLOPPY (w) DISC? OR DISKETTE?)

MULTILINGUAL FACILITIES

1. Query Input in User Language
2. Query Formulation and Modification in User Language
3. Screen Messages in User Language
4. Automatic Translation of Search Terms into Language of Database

MODELLING THE USER

1. CAPABILITIES
   - Experience in Searching
   - Knowledge of Subject
   - Knowledge of Languages
2. WORK CONTENT
   - Practical (Engineer)
   - Research
   - Administration
   - Student
   - Amateur Interest

   etc
3. PREFERENCES IN CURRENT SESSION
   - Type of Information
   - Volume of Output
   - Precise or Broad

CLARIFYING A QUERY

The Use of Frames and Semantic Categories

QUERY TERM: Pruning

CATEGORY: Operation

ASSOCIATED CATEGORIES: Plant:?
Tool:?

Fill Empty Slots by Questioning User

HELP IN SELECTING DATABASES

Query Input

Look up Terms in Dictionary
Find their Class Numbers

Climb Classification to Reach Subjects used in Database Index

Select Database/s

Refine Selection
Depending on:

1 Host
(Does user have password?)

2 Type of Information
(Is it included in database?)

3 Language of Database
(Can user understand this language?)

Offer Final Selection to User for Confirmation

BRIDGING THE GAP

1. Helping User to Select Sources of Information
2. Accepting Queries in the User's Language
3. Intelligently Processing Queries to Form an Optimal Search Statement
4. Automatically Putting the Search into the Form Required by the Online Host
5. Handling Automatically all Telecommunications with the Host

HELP DURING SEARCH

1. Initial Search Strategy
   - Dictionary supplies relations between search terms
   - System inserts Boolean and adjacency operators and uses truncation
2. Translation for Host
   - System uses host command language and transmits search
3. Search Modification
   - Thesaurus offers user broader, narrower and related terms

MULTILINGUAL FACILITIES

1. Screen Messages in User Language
2. Query Input in User Language
3. Query Formulation and Modification in User Language
4. Database Selection Takes into Account which Languages User Can Read
5. Automatic Translation of Search Terms into Language of Database Index

SOURCE DATA DICTIONARY

This will Store Data about Hosts and Databases

- Command Languages
- Type of Database
- Searchable Fields
- Language of Database Records
- Document and Information Types Included
- Character Sets

Database Publishers' Challenges for the Future

By
Barbara Lawrence
Administrator
American Institute of Aeronautics and Astronautics
Technical Information Service
555 West 57th Street
Suite 1200
New York, NY 10019

INTRODUCTION

As decades and centuries reach their turning points it is useful to pause a moment and think about the future. Now is one of those times, and it can be a bit frightening for database publishers. There are so many competing forces, technologies, and user communities that it is hard to forge our own vision.

This analysis of the path to the future tries to keep the focus on the role of database publishers. As an aid to achieving this focus, this paper takes as its theme some recent words from Everett Brenner, long a gadfly of the database and secondary information service community.

> *"The success of information systems using computers, whether in research or business areas of the corporation, rarely ever depends on the technology. Success finally hinges on the information itself – its quality and ability for humans to communicate real needs which can be translated into a digestible and retrievable form."*[1]

With the focus clearly on our role as information providers who are concerned with providing information content through various access channels, we can evaluate the trends and the challenges. The following discussion will consider briefly users, who are the reason for building databases, the very enticing visions of how databases can be used, and finally the more specific challenges that database producers must meet in order to remain effective in these future scenarios.

AEROSPACE SCIENTISTS AND ENGINEERS

There are many different information-using communities, and the differences between them may be large, even within categories such as R&D or engineering. Thus, in order to meet the challenges ahead as database publishers, it is necessary first to understand the particular nature of the specific users. It is important to reduce communications barriers so that we, the information providers, may do our part in aiding aerospace scientists and engineers to do their best work.

The NASA/DOD Knowledge Diffusion Project [2] is providing an up-to-date baseline of information habits of aerospace scientists and engineers. This kind of understanding is critical. Without it, information professionals often define user needs in terms of information systems tools, rather than as the need to solve a technical problem. For instance, one study [3] defines needs in engineering as including easier access to databases and computer program libraries. However, these are things that the information provider must do to meet the real need, which is for information, not systems.

The following is a generalization from the observations that have been made at AIAA's Technical Information Service about the aerospace scientist and engineer as represented by the AIAA member.

Engineers are trained to be "do-it-yourselfers." Coupled with the large scale of many aerospace and defense projects, one result is the "not-invented-here (NIH) syndrome," the tendency to

believe that all of the required expertise is located within the home organization. It is especially prevalent in the aerospace community. Recent National Science Board studies find that US scientists and engineers continue to overcite the US literature and to cite primarily their own specific field [4]. They seem somewhat unadventurous in exploring the literature, even as they engage in very adventurous technology programs.

This insular thinking is so pervasive, that the NIH syndrome even appears in the field of space law. One would expect that dealing with space policy would be inherently international.

Yet the picture should not be painted as overly bleak. Although the focus for aerospace scientists and engineers is on known sources, and the starting point for information location is collegial sources, aerospace scientists and engineers do use the literature, particularly journal articles, conference papers, and reports. Pinelli found that ninety percent indicate that technical communication is very important. [5]

The scope of aerospace and defense is multidisciplinary and older literature is as important as the latest leading edge material. It may well be that aeronautics libraries, if forced to choose one collection only, would keep the old National Advisory Committee on Aeronautics (NACA) reports first. This is a body of high quality, well documented knowledge from 1915-1958. Thus, the original quality of the input to the database really matters. Perhaps this focus on reliability should be used in analyzing other databases. Quality is a hallmark of building the database at AIAA, but maybe it is time to test perceptions.

Aerospace scientists and engineers are not merely looking for information; they are trying to solve problems or find specific answers. It is therefore necessary to organize high quality information in an accurate and effective fashion and then to design the tools, the technology, for ready access.

That aerospace scientists and engineers are not intimidated by information technology is supported by Pinelli's findings [6] that these scientists and engineers understand what the technologies are and that a majority even identify the following technologies as "I don't use it, but may in the future":

- laser disk/video disk/compact disk
- video conferencing
- electronic bulletin boards
- electronic networks

These findings are supported by unpublished AIAA member surveys, which show extensive use of PCs for external network communications and information retrieval, and a strong interest in information-technology-based technical information services. These scientists and engineers are also at ease with related terminology, such as electronic networks, computer conferencing, and optical disk.

One reason to give scientists and engineers credit for being comfortable with information technology is that they are the ones who built it. It is the job of the information professional to organize the information content, but this should be done in collaboration with the users. Scientists and engineers are not unaware of the problem of organizing and making accessible large volumes of information. In 1984, a NASA Ames scientist [7] described a vision of networks and workstations for instant comprehensive access in a paper about "Information Systems: Issues in Global Habitability". On this all important issue, the authors noted that "little has been said regarding the problems of how to collect, access, manipulate and store data on such an overwhelming scale." This large scale of scientific and technical information is a particular issue in aerospace, but is not unique for these databases.

There is now the opportunity to work closely with the constituency. If database producers treat their customers as partners, they will work with them. In this way, by reinforcing our appreciation

for the information requirement of users, database producers will deliver the most appropriate information and avoid being carried away by what may excite us the most.

THE FUTURE - THE VISIONS

Database publishers have a variety of visions or scenarios to use as a planning basis. These visions generally have two common elements: desktop access (and note, no intermediary) and virtual integration (any type of information available through a single interface).

The parent of these futures for scientific and technical information was, Vannevar Bush, a US science advisor, who coined the phrase "MEMEX machine" in 1945. [8] In those times he saw the data store as microfilm but the data structure as hypertext, all integrated with the desktop (or actually, the desk).

Another early vision was Organization X. [9] This was to be a coordinated effort on the part of the scientific and technical information abstracting and indexing services to reduce overlap and expand coverage through a joint venture operation, which would then facilitate project-focused repackaging.

However the NFSAIS members (National Federation of Science Abstracting and Indexing Services, now known as the National Federation of Abstracting and Information Services/NFAIS) who commissioned the study which recommended Organization X, were overwhelmed with concerns about competition, and no grand schemes were undertaken. Fortunately, NFAIS is currently working on the themes of cooperation and standards, working to provide leadership to database publishers.

In the early eighties, James J. Harford, the AIAA Executive Director, articulated a vision of a global science and technology network (STN) [10] to serve five million scientists and engineers from their desktops with all types of information: journal literature, conference proceedings, design specifications, meeting programs, electronic mail, handbook data; all internationally available with language translation, news, discussion groups, and so on.

Harford wrote that there are "no technological show-stoppers in the way of realization of this scenario. There are formidable costs and difficult organizational and political barriers." He then described his experiences trying to get the engineering societies to cooperate to achieve this vision. Harford was no more successful than NFSAIS was in joining abstracting and indexing services. But the American Chemical Society took on the STN notion in their own way.

A key lesson learned from reviewing these visions is the need for a truly cooperative team, if the information and the technology are ever to come together. In August 1990 the Southern California Online Users Group (SCOUG) met to analyze database quality criteria. They developed a quality rating scheme that considers that the database and the system must be evaluated together; after all they are only useful together.

More recent versions of the "vision" begin to address the effort that it will take to get there. Martha Williams understood that it will take extensive research and development to build these transparent information systems.[11] Are database publishers contributing sufficiently to this effort? The commercial publishers perform market research and product development. In the US, the National Science Foundation no longer funds academic information science research, supporting computer science instead. How will this research be done? The NATO Advisory Group for Aerospace Research and Development, Technical Information Panel is currently developing a research agenda. Other information professional organizations, such as the Special Libraries Association and Medical Libraries Association, also recognizing the need for research, have developed research agendas too. It is hoped that these agendas will encourage both researchers and funders.

Mick O'Leary expresses the vision and the challenge another way, predicting that the 1990s will be the Age of Access. He means by this that "the 90s will be less concerned with seeking new information than with organizing, distributing, and applying that which we already have." [12]

Monitor, arguably the most insightful, but surely the most enjoyable reading for the commercial/external database community, tried to foretell the '90s. They believe that the message of the '80s was that not everyone needs access to huge databases and that more is not always better. They predict that the '90s, fostered by the power of the desktop workstation, will see integration of internal and external data. Database publishers, serving the professional community, will see a shift from print information to electronic "in either ASCII or digitized form and ... thus more easily exploitable at the local workstation." [13]

Vast, resource-rich networks are moving from vision to reality. One of the major challenges for database producers is to relate to these networks. Network developers have their work to do. For instance NASA has five networks, not all of which interconnect. To solve this dilemma, companies have formed (e.g., DASNET and GEONET) to switch messages among electronic mail systems. And in the US, it is estimated that the largest network, the INTERNET, consists of 330,000 computers and 2 million users. The proposed National Research and Education Network (NREN) will expand this INTERNET and is the beginning of a "focus on networks as part of an infrastructure for access to information resources, rather than simply as highways providing connectivity and carrying bits from one place to another." [14]

A leading force behind the NREN, Robert Kahn, takes the "wired nation" vision the next step, saying that "you'd like all participants in countries to share information in a uniform way." [15] Kahn, the designer of ARPANET (the first network), recognizes the need to deal with issues beyond moving megabytes of data around and that data organization, standards, intellectual property rights, data access rights, are all issues that we must address. The dialog on these topics should include a significant presence from the database producing community, although it does not yet do so. NFAIS has opened discussions with Kahn as a first step in this direction.

The urgency of the issues mentioned above is emphasized when we acknowledge the interface part of Kahn's vision. He is building bits of artificial intelligence software called Knowbots, which will take the query and find information anywhere in the network in a process that is completely transparent to the user. So records and even parts of records will be disconnected from the databases and host systems. Rather than ignoring the challenges posed by this image, database publishers must step up to them. Users will appreciate it.

Networks already have demonstrated their impact. Recently the international mathematics community used electronic mail networks to solve a problem in a matter of weeks, as competing teams, aided by the ease of communication and the energy of competition, rushed to solve a complex problem. One mathematician said that "computer networks are replacing professional journals and conferences." [16]

One disquieting facet of these visions of instant access to a universe of information is that the visionaries tend to ignore the role of the professional database publisher. Yet here are the information scientists who can address the problems of information structure for interchange and access. Here are the people who manage the quality of the data. If information access is to be meaningful, the information "must be complete, current, and presented in ways that encourage its use. It must also be reliable, accurate, and absolutely dependable." [17] Lois Granick stated the charge clearly, that database publishers, must take an active leadership role or lose control and, in so doing, the world loses an opportunity to become an "informed society."

FROM VISION TO REALITY — SEGMENTATION

Although all of these visions express the notion of integration of multiple resources, the reality, pushed by the same technological opportunities, is really diversity, at least from the database

publisher perspective. Morry Goldstein, President of IAC, says that "the wide variety of ways user organizations wish to use the information we offer for sale means we are being forced into an RFP (request for proposal) kind of business wherein we respond to very specific user requirements with a proposal tailored to that environment." [18]

The user community is not monolithic. It is not even as simple as intermediaries or information professionals and end-users. One can subdivide these groups, for instance, reference librarians and literature searchers, special librarians and academic research librarians, engineers and scientists, researchers and managers. The categories can and should be expanded, if not for the purpose of developing different databases for each, at least for analyzing how the database may be used by each segment. Database publishers have so far tailored products for professional, library, management, and consumer communities. In the future, more differentiation may occur.

Within the aerospace STI-using community, the user environment is also segmenting, with different technology platforms being favored for information access by different groups. The academic community has adopted CD-ROM and, in large universities, is moving to "local online" for widely used databases. (local online is online from a university mainframe for institution-wide access, usually combined with the online library catalog). Government users are beginning to rely on networks. Corporate users may go "local online" but are also becoming bulk buyers from the vendors' services, because they generally use so many databases.

While most online use in aerospace is mediated by the information professionals, the scientists and engineers, at least those who are AIAA members, tell us that they want to search directly from their own desktop. AIAA is attempting to meet this challenge of diversifying user segments with a new service, a menu-driven front-end on DIALOG, called the AIAA Connection. The AIAA Connection will do more than just search the Aerospace Database. It will also provide electronic mail, bulletin boards (for member committees, for AIAA headquarters departments, for meeting schedules, etc.), document ordering, and, ultimately, other databases.

Meanwhile, many aerospace scientists and engineers are using optical storage for raw data. Also, aerospace corporations are evaluating optical storage for archives. If these platforms become wide-spread within the end-user base, should database publishers redesign their products for end-users in localized settings?

World political changes will also have an impact on aerospace scientific and technical information service. For now it appears that the changes will be negative, resulting in closed libraries, cost-recovery requirements for the Department of Defense Defense Technical Information Center, and information budget cuts generally. Yet at the same time, these changes seem to be encouraging research and development, the major process for scientific and technical information use and generation. In fact, AIAA membership is growing, not declining, as some might expect in these confused times. So the information providers will be even more pressed in meeting the needs of the many segmented users and environments requiring our services. It will be necessary to be creative, cooperative, and flexible.

CHALLENGES FOR DATABASE PUBLISHERS

Flexibility is the critical keyword for the future of database publishers. Some other good keywords are progressive, technology-driven, strategic planning, and user-focused. But most important is, *flexibility*.

In 1985, Martha Williams, a longtime observer of the database world, identified the issues relating to electronic databases as "public-private sector competition, transborder dataflow, copyright, downloading, and the changing roles in database generation and processing." [19] These issues are still around, but, today's challenges have a lot more to do with leadership and determination.

Getting Digital

If the future is electronic distribution, then the first challenge, and the key to flexibility, is to digitize the data – all of it. This is not so simple. First publishers must move beyond bibliographic information and on to full-text, numeric, and image information. Database publishers will have to manage and integrate a variety of input techniques, such as manual data entry, data conversion contractors, scanning, and author electronic submission. [20] We will need to avoid seduction by enthusiastic sales people, because selecting the methods that fit any particular material is not necessarily easy. There are all sorts of details to be resolved. Some examples: Scanning from thin paper is a problem, because back printing bleeds through; Authors do not like to code manuscripts; Data conversion contractors are fast and low price, but cannot correct errors; Who has the right to correct another publisher's text?

However, managing information means that database producers must do more than create an electronic store. The EUSIDIC community, in the spring of 1990, focused the challenge on data structure. [21] To do this, database publishers need to be willing to cross barriers, even to work with other publishers. How else can we use hypertext to go from data tags to the numeric tabular data, from the role indicator to the chemical reaction, from the description to the computer code, or from the abstract to the full-text? How else can we move beyond Boolean logic and controlled vocabulary to the world of Robert Kahn's Knowbots? How else can we move from the retrieval of information to enable the use and manipulation of that information? And how else will we separate quality from quantity?

Database producers will have to change their practices and strategies to meet the challenges of creating new data structures for electronic dissemination. We must support real information science research. And , even more difficult , we must develop a cooperative effort and a willingness to commit to using standard practices.

Database publishers once rejected the notion of Organization X. Today they hardly use the standards that do exist (ISO, NISO). Will they take the primary publishers as an example and adopt a Standardized General Markup Language (SGML) application for databases? (This positions publishers for acceptance in many user environments, since database loading will be consistent). Will they control the future, or let others create the standards, as database vendors have done in the last decade? If the database professionals sit back, the network developers will lead, without benefit of our understanding of information science, database structure, information retrieval, and so forth.

Quality

Creating a complete electronic data store is not the whole answer unless the challenge of quality is met. Quality is this year's hot topic, at least in the US, and with good reason. We face a less passive audience, one that can select from a variety of resources. Expert users have become increasingly vocal. As mentioned earlier, SCOUG is developing a rating scheme for database quality, one that goes to considerable, specific depth.

End users are clear that they want quality not quantity. AIAA members often say that their library gives them too much information, rather than the answer to their problem.

Total Quality Management, and its related concept of continuous improvement, has become entrenched in government contracting within the US Department Of Defense and the National Aeronautics and Space Administration. It should apply to information as well as manufacturing.

Aitchison raised the issue of quality in his 1988 Miles Conrad Memorial Lecture, "Aspects of Quality." [22] Several recent articles have identified certain quality characteristics in common, as shown in the following table.

QUALITY CHARACTERISTICS FOR DATABASES

	Aitchison[23]	Potzscher[24]	Maxon-Dadd[25]	Mintz[26]	SCOUG[27]
Accuracy	X		X	X	X
Reliability	X		X	X	X
Consistency	X	X	X		X
Comprehensive	X	X	X		X
Timely	X	X	X	X	X
Cost/Value		X			X
Support				X	X
Accessible				X	X

Additional characteristics include integration (with the retrieval system, links to full text), controlled vocabulary, and error correction policy. The list is long. It is important that database creators work with users to establish a thoughtful evaluation process and respond to constructive critiques of their products

Business Issues

After the objective of creating digital databases is achieved, we next face a very complex set of business challenges.

Pricing is perhaps the primary conundrum, because if we are wrong, we may not receive sufficient revenue to continue to produce the database. While for most there will still be print products, electronic dissemination in multiple environments will begin to dominate. There is a cost in delivering for multimedia, segmented user communities, and in achieving a new quality standard. The intellectual efforts, such as for abstract writing, still have a price tag. It is not a trivial challenge to learn how to price according to the value of the information, but we will have to try.

Packaging is an issue quite related to pricing. Delivering to some of these diverse distribution channels may require incorporation of hardware or software with the database. Many database publishers have never had sufficient information technology-skilled staff to manage the design, much less the customer service aspects, of such packages.

The third business challenge is to create fair business agreements between data creators, distributors, and users. The balance between having a crystal clear and comprehensive license agreement and retaining some sense of trust and respect among the partners is very difficult. In multinational arenas the legal issues are surely more complex. We cannot be naive, but we should not be controlled by our attorneys. Ron Dunn, now of Maxwell Multimedia Publishing Group, always said, and I paraphrase, that lawyers are there to get you out of trouble, after the fact, and not to keep you out. Or else you would never do anything in the first place.

Policy Challenges

The business issues are real ones, critical to our survival. The policy challenges are equally serious. One long-standing topic, a hot one in government-dominated aerospace and defense, is the role of the public sector as information provider. The US Congress sponsored a recent study, Helping America Compete [28] which urges the federal government to recognize the value of scientific and technical information and to reinforce its agency and unified efforts. In Europe DG-XIII of the EEC has specifically and strategically supported information service and information technology developments. In Japan the approach is to organize government run information collection and dissemination centers. Overall, the private sector has benefitted from government support and has returned the service by adding value and dissemination expertise to basic government data.

However, at least in the US, there are tensions, as government begins to add value to its databases and, in particular, becomes CD-ROM crazed. Another source of tension can be regulatory, a reality as AT&T and the Baby Bells push for the ability to provide information services.

The push-pull of this relationship can only be managed with honest and open dialog. If we play power games, we will be distracted from the information providing objective.

Addressing public-private sector relationships, is part of the process of readdressing all of our relationships. Networking, for example, may alter the roles of all of the players: database publishers, authors, primary publishers, vendor-host, gateway, library, etc. One way to approach this is a pending effort by NFAIS to revisit their Gateway Code of Practice, a matrix of rights and responsibilities, in light of the diversification of online environments to include networks, local online, and local area networks.

Not to be lost in revisiting the relationships in the information creation and delivery chain are intellectual property rights. This topic is at least a full paper in itself. Intellectual property is an active arena, and some current topics are:

- Database deposit, US
- Copyright of databases, EC
- Author awareness of rights
- Role of reproductive rights organizations
- Chain of rights (author-primary publisher-full text database-online service)
- Electrocopying

The last group of challenges - ethics and cooperation – have been written of before. There is an urgency incumbent upon us now to meet these challenges at a high level if we are to achieve our mission in the complex future environment. It is only by taking the high moral road that we will be able to handle the other challenges.

Weil [29] and Granick [30] have been our most consistent spokespersons on the subject. Let us take the time to reflect, to remember why we are here, and to review our mission – our focus. Most in the scientific and technical information profession are here because of a strong belief in its value in the world. If this is still true, we must recognize our purpose and use it to empower us to address our roles and responsibilities with seriousness.

CONCLUSION

Aitchison said it so well:
"Instead of worrying too much about who's doing what to whom, we should get on with our work and keep developing our plans without looking over our shoulders all the time. In particular, we should resolve that if anyone ever gives our users what they really want, we shall be part of the team that does it." [31] If we do this in a cooperative and proactive manner, we can achieve our goals.

This paper discussed the challenges for data content and electronic distribution, quality and standards, flexibility and leadership, ethics and cooperation. If we see our job as delivering the right information at the right time and place, then we must get on with meeting these challenges.

REFERENCES

[1]Brenner, Everett, "Executive Information Systems, Part III," Monitor, Number 113, July 1990, pg. 9.

[2]The NASA/DOD Knowledge Diffusion Research Project: A Research Agenda, Principle Investigators Kennedy, John M., Center for Survey Research, Indiana University and Pinelli, Thomas E., NASA Langley Research Center.

[3]Charles Goldstein, "ASIS 2000 - Compter Science and Engineering," Bulletin of the American Society for Information Science, October/November 1989, pg. 15.

[4]National Science Board, Science & Engineering Indicators, 1989. Washington, DC: U.S. Government Printing Office, (NSB 89-1), 1989.

[5]Pinelli, T.E., Glassman, M., Oliu, W.E., and Barklay, R.O., Technical Communications in Aeronautics: Results of an Exploratory Study, NASA TM 101534, Part I, NASA 1989.

[6]Pinelli, T.E., et. al. loc. cit.

[7]Norman, S.D., Brass, J.A., Jones, H., and Morse, D.R., "Information Systems: Issues in Global Habitability," AIAA-84-189, AIAA 22nd Aerospace Sciences Meeting, Reno, Nevada, January 9-12, 1984.

[8]Bush, Vannevar, Science The Endless Frontier, Washington, DC, U.S.G.P.O, 1945.

[9]Robert Heller and Associates, A National Plan for Abstracting and Indexing Services, NFAIS, 1963.

[10]Harford, James J. and Lawrence, Barbara, "Future Patterns of Scientific and Technological Information Exchange: A Scenario," AAAS Annual Meeting, New York, May 29, 1984.

[11]Williams, Martha E., "Transparent Information Systems Through Gateways, Front Ends, Intermediaries, and Interfaces," Journal of the American Society for Information Science, 37(4): pp. 204-214, 1986.

[12]O'Leary, Mick, "Databases of the Nineties: The Age of Access," Database, pp. 15-21, April 1990.

[13]"In Six Weeks Time It Will Be 1990...," Monitor, Number 105, November 1989.

[14]Lynch, Clifford, A., "The Growth of Computer Networks: A Status Report," Bulletin of the American Society for Information Science, pg. 11, June/July 1990.

[15]Richards, Evelyn, "The Data Deluge: Exotic Electronic Systems May Hold Key To Future Access," The Washington Post, Vol. 112, No. 293, Sect. H., Sept. 24, 1989, pg. 1.

[16]Kolata, Gina, "In A Frenzy, Math Enters The Age Of Electronic Mail," New York Times, June 26, 1990.

[17]Granick, Lois, "Laying the Foundation for An Informed Society: A Critical Role for the Secondary Information Services," Miles Conrad Memorial Lecture, NFAIS Newsletter, 32(4), April 1990, pp. 41-49.

[18]Hogan, Tom, "Database Producers Should Prepare For 'Radical New Thinking,' says IAC's Goldstein," Information Today, Vol. 7, No. 5, May 1990.

[19]Williams, Martha E., "Electronic Databases," Science, Vol. 228, April 29, 1985, pp. 445-456.

[20]O'Leary, Mick, "Producing A Database: Many Choices For Data Entry," Database, Feb. 1990, pp. 38-40.

[21]Monitor, "The Problem of Access and Data Structure," No. 110, April 1990, pp. 10-11.

[22]Aitchison, T.M., "Aspects of Quality," Miles Conrad Memorial Lecture, NFAIS Newsletter, 30(2), April 1988.

[23]Aitchison, T.M., op. cit..

[24]Potzscher, Gunter, and Wilson, A.J.C., "User Needs in Chemical Information," J. Chem. Inf. Comput. Sci., (30), 1990, pp. 169-173.

[25]Maxon-Dadd, Josephine, "Refurbishing An Elegant Victorian Database - A View from DIALOG," Trends in Database Design and Customer Services, ed. Schipper, W. and Unruh, B., NFAIS, Philadelphia, 1990.

[26]Mintz, Anne P., "Quality Control and The Zen of Database Production," to be published November 1990, Online.

[27]Southern California Online Users Group, Retreat IV, Measuring the Quality of Data, August 1990.

[28]U.S. Congress, Office of Technology Assessment, Helping America Compete: The Role of Federal Scientific and Technical Information OTA-CIT-454, Washington, DC, U.S. Government Printing Office, 1990.

[29]Weil, Ben H., "Information Transfer in A Time of Transition: The Need for Community, Organizational, and Individual Empathy and Ethics," ed. Neufeld, M.L., Cornog, M., Speer, I.L., Abstracting and Indexing Services in Perspective, Miles Conrad Memorial Lectures, Information Resources Press, 1983.

[30]Granick, Lois, "Proposed Code of Ethics For The Information Community," ed. Neufeld, M.L., Cornog, M., Speer, I.L., Abstracting and Indexing Services in Perspective, Miles Conrad Memorial Lectures, Information Resources Press, 1983.

[31]Aitchison, T.M., "The Database Producer In The Information Chain," Journal of Information Science, 14, 1988, pp. 319-327.

EXEMPLES DE BASES DE DONNEES UTILISANT DES TEXTES NUMERISES (TEXTES INTEGRAUX OU RESUMES)

par

Pascal Pellegrini et Philippe Laval
CORA SA
93, avenue de Fontainebleau
94270 Le Kremlin-Bicêtre
France

Le but de cet exposé est de montrer les avantages d'une base documentaire indexée directement sur le texte, par opposition aux systèmes classiques à thésaurus.

Cet exposé est divisé en trois parties : après un rappel de l'état de l'art dans le domaine de la linguistique informatique, nous expliquerons comment DARWIN™ réalise cette indexation. Enfin, nous conclurons avec plusieurs exemples d'applications.

I. La linguistique informatique.

Le but de la linguistique informatique est de doter l'ordinateur de capacités de traitement de la langue naturelle incluant un minimum de compréhension. Cette définition exclut en particulier toutes les applications de type traitement de texte, traitement statistique ou recherche simple de mots clefs à partir d'un thésaurus construit manuellement.

Dans ce chapitre, nous ne nous intéresserons qu'à la langue écrite, une fois tous les problèmes de saisie résolus.

I.1 Le domaine

I.1.a Les outils

- Parseurs (ou analyseurs syntaxiques): il s'agit de programmes qui calculent la représentation syntaxique d'une phrase.
- Grammaires: ensemble de règles permettant de décrire les constructions syntaxiques autorisées.
- Lexiques informatiques: citons le système DELA du LADL (Université de Paris 7) qui couvre plus de 700 000 formes fléchies des mots français et 400 000 mots composés de la langue française,
- Outil de représentation des connaissances (pour le traitement sémantique des textes).
- Générateur de langue naturelle.

I.1.b Les applications.

- Traduction assistée par ordinateur.
- Aide à la rédaction : correcteur orthographique, grammatical ou stylistique, dictionnaire de synonymes...
- Recherche documentaire.
- Interface en langage naturel.

I.1.c Les réalisations industrielles.

- Interrogation de base de données:
  - Pages jaunes de l'annuaire (GSI-Erli)
  - Q&A (Symantec)
  - Clout

- Vérificateurs orthographiques :
  - Writer Workbench (ATT)
  - Alpha (Borland)
  - Sprint (Borland)

- Traduction assistée :
  - SysTran (Gachot)
  - TAO (Alp Systems)
  - Weidner
  - Atlas (Fujitsu)
  - Pivot (Nec)

- Recherche documentaire :
  - Baris (Batelle)
  - Spirit (Systex)
  - Darwin (Cora)

I.2 Le problème de la compréhension.

I.2.a Les trois niveaux d'analyses de la langue écrite.

Il est généralement reconnu qu'il existe trois niveaux d'analyses pour la langue écrite, à savoir:

- le niveau syntaxique, qui décrit les règles de formations des phrases et les transformations possibles d'une structure en une autre (par exemple : passif vs. actif, affirmatif vs. négatif).
- le niveau sémantique, qui traite du sens des mots et des moyens de calculer le sens d'une phrase à partir du sens de ses mots.
- le niveau pragmatique ou extra-linguistique qui consiste en la connaissance du monde environnant.

I.2.b La compréhension.

le but de la compréhension automatique du langage est de transformer une phrase en une structure interne explicitant les relations entre les différents concepts de la phrase, sachant que:

- la représentation doit faciliter les inférences possibles,
- le degré de compréhension dépend de l'objectif visé,
- il n'y a pas isomorphisme entre la phrase et la structure interne.

I.2.c Les principales difficultés

Nous allons dresser une sorte de catalogue des principales difficultés liées à l'analyse automatique du langage :

- Les ambiguïtés lexicales (homographie):

  La petite brise la glace.

  La belle ferme la porte..

  Time flies like an arrow.

- Les polysémies:

  voler (en l'air, un objet)

  arm, beam

- Les attachements prépositionnels:

  Elle mange le poisson avec des arêtes.

  Elle mange le poisson avec une fourchette.

  He saw the cat with a long tail.

  He saw the cat with a telescope.

- La portée des conjonctions:

  Les chats et les chiens sans queue.

  Les poulets et les chiens sans queue.

  Cats and dogs without a tail.

  Chickens and dogs without a tail.

- Les tournures elliptiques:

  Quelle est la capitale du Guatemala ?

  Du Mexique ?

  Which is the capital of Guatemala ?

  Of Mexico ?

- Les anaphores:

  Le professeur a envoyé le cancre chez le censeur car :

  1. il en avait marre

  2. il voulait le voir

  3. il lançait des boulettes

- Vocabulaire incomplet.
  - Néologismes.
  - Créations d'auteur.

- Les expressions figées et semi-figées:

  Prendre la poudre d'escampette

  à couteaux tirés

  raining cats and dogs

  to have to see a man about a dog

- Structure syntaxique identique avec un sens très différents:

  Jean est difficile à convaincre.

  Jean est habile à convaincre.

  John is easy to persuade.

  John is eager to persuade.

- Structures syntaxiques différentes avec sens proches:

  Elle marche en souplesse.

  Sa démarche est souple.

  She walks in a relaxed way.

  Her walk is relaxed.

- Sous-entendus (compris par tout humain):
  J'ai été dans trois librairies ce matin. (je n'étais pas dans les trois librairies à la fois)

  I had been in three bookstores this morning when it began to rain.

- Inexistence d'un ensemble de règles qui engendreraient toutes les phrases acceptables d'une langue et rien que celles-là: de nombreuses connaissances sont inexprimables sous forme de règles (pour les résolutions pronominales par exemple).

- Imperfection des textes (fautes de frappe, d'orthographe, de grammaire, de ponctuation, de style).

- Manque d'intérêt des linguistes purs pour les textes réels.

II. Les bases de données constituées de textes intégraux

Dans de nombreux domaines (finance, assurances, administrations, journaux...) les données sont présentées sous forme de textes écrits. L'un des problèmes de l'intelligence artificielle est la représentation des connaissances et la conversion de ces dernières en des formalisme capables d'être interprétés par des ordinateurs.

Dans la suite, nous ne nous intéresserons pas au problème de la compréhension et du traitement des textes. Par contre, nous allons nous attacher tout particulièrement aux méthodes à mettre en œuvre pour être à même de pouvoir retrouver toutes les informations contenues dans ces textes.

Deux méthodes différentes sont utilisées pour la recherche documentaire: les systèmes à thésaurus (ex: BASIS) et les systèmes indexant directement le texte (DARWIN).

II.1 Les systèmes à thésaurus

Ces systèmes ne font pas intervenir de traitement de la langue. Pour chaque famille de textes, un thésaurus est constitué manuellement. Les textes sont indexés sur les mots contenus dans le thésaurus. Cette méthode est facile à mettre en œuvre de manière informatique. Cependant:

- elle ne garantit pas que toutes les informations contenues dans les textes pourront être retrouvées: l'oubli d'un mot-clef entraîne l'impossibilité de retrouver les portions de textes les contenant. La constitution du thésaurus est donc une étape délicate demandant beaucoup de travail et une connaissance parfaite des textes à indexer,

- la mise à jour du thésaurus (ajout d'un nouveau mot-clef par exemple) entraîne une réindexation de toute la base: ceci peut être gênant si la base est très importante,

- la modification de la base de données documentaire peut entraîner une mise à jour du thésaurus (pour prendre en compte un nouveau domaine par exemple). Le coût de la maintenance est donc important,

- l'indexation mot à mot introduit des bruits lors des recherches. Par exemple, les mots "**or**" et "**car**" posent problème: s'ils sont mis comme mots-clefs, ils introduisent un bruit considérable, car tous les **or** et **car** du texte sont indexés, même si ce sont des conjonctions de coordination. Par contre s'ils ne sont pas mis comme mots-clefs, le "silence" risque d'être important. De façon similaire les homographies (ex: la joue/il joue) ne sont pas levées. Ce bruit est encore plus gênant en anglais où quasiment tous les mots sont des homographes nom/verbe. Voici quelques exemples de bruit:

  ***Or**, il faisait chaud.*

  *La petite Marie **joue** dehors.*

  *Les poules de Marie **couvent**.*

- la recherche de **conseil général** va donner comme réponse tout ce qui touche **conseil** et tout ce qui touche **général**, ce qui introduit encore un bruit considérable. Ce problème est partiellement résolu dans certains systèmes grâce à la notion de proximité: dans notre exemple il ne faudra fournir de réponses que si **général** suit directement **conseil**, ce qui fournit les bonnes réponses.

Afin de garantir une plus grande fiabilité et un coût moindre, nous voyons qu'il est nécessaire de supprimer la constitution du thésaurus; c'est ce que réalisent les systèmes indexant directement les textes.

II.2 Les systèmes indexant directement les textes

DARWIN part d'une approche radicalement différente: plutôt que de constituer un thésaurus manuellement, le texte est analysé de façon à identifier les concepts (groupes nominaux étendus) qu'il contient. L'indexation se fait sur ces concepts. Cette indexation se fait à partir d'une analyse syntaxique appuyée par des règles linguistiques. Par exemple, le texte:

Le petit chat est noir.

sera indexé sur les concepts "**petit chat**" et "**noir**". Cet exemple est bien entendu très simple et DARWIN est capable d'analyser tous les textes ayant une orthographe et une syntaxe correcte.

Cette analyse est:

- automatique: elle ne nécessite aucune intervention particulière,

- indépendante du domaine dans lequel on évolue, puisque la seule contrainte est que les textes respectent les règles de la grammaire. Cela signifie en particulier qu'il n'est pas nécessaire d'apprendre à DARWIN les expressions significatives du contexte (qu'il s'agisse de textes juridiques, techniques, littéraires...),

- fidèle et complète: les expressions retenues constituent le fond documentaire, comme pour les systèmes à thésaurus. Cependant l'indexation automatique repère tous les objets sans exclusion et avec une très grande fiabilité, alors que les systèmes à mots-clefs sont souvent peu fidèles et dépendants des personnes qui indexent:
  - les bruits dûs aux homographies nom/verbe sont supprimés puisque l'analyse linguistique permet de détecter les verbes,
  - les bruits dûs aux homographies dûes aux mots grammaticaux (comme **or** et **car**) sont eux aussi éliminés grâce à l'analyse du texte,
  - les problèmes de proximité disparaissent: la recherche de **conseil général** donnera comme réponse tous les groupes nominaux contenant effectivement **conseil général** et non ce qui touche **conseil** et **général**.

DARWIN permet d'organiser les textes de manière arborescente en documents, sous-documents etc... afin de structurer les textes.

II.2.a Constitution des bases textuelles

Les étapes suivantes sont nécessaires:

- introduction des textes, soit en les tapant directement, soit à l'aide de scanners,
- correction des textes introduits afin qu'ils respectent les règles de la grammaire. En effet, les scanners introduisent généralement des erreurs (sur par exemple),
- structuration des textes (organisation en documents, sous-documents...),
- indexation des textes structurés.

i. Introduction des textes

Trois cas se présentent:

- le texte existe déjà sous forme informatique; il a a priori un format quelconque (Word, WordStar, Sprint, WordPerfect...). Il faut donc le convertir en ASCII. Cette opération ne pose pas de problème particulier puisque tous les traitements de texte proposent une sortie ASCII étendue,
- le texte existe uniquement sous forme papier; il est numérisé avec un scanner. Cette opération nécessite une relecture du texte, car certains caractères sont souvent mal reconnus. Les erreurs les plus fréquentes se font entre les **i** et les **l**, les **1** et les **I**, les **8** et les **g**. Il faut noter que même s'il n'y a qu'une erreur dans le texte, il faut relire tout le texte,
- le texte n'existe pas encore; il est alors saisi directement sous DARWIN.

ii. Correction des textes

Cette étape a pour but de corriger les erreurs résiduelles dans le texte à l'aide de correcteurs orthographiques. Ces erreurs sont essentiellement de deux types:

- les erreurs de saisie, et celles induites par les scanners: inversion de deux caractères, oubli d'un caractère, erreur de frappe sur un caractère, caractère dédoublé. Ce sont des erreurs lexicales.
- les phrases mal construites, qui ne sont pas correctes grammaticalement. Ce sont des erreurs syntaxiques.

1 Erreurs lexicales

Elles sont assez facilement détectables en comparant chaque mot du texte avec un lexique. Plus le lexique contiendra d'entrées, meilleure sera la correction. Cependant, le nombre de formes fléchies d'une langue (plus de 700 000 pour le français) fait qu'il n'est pas possible de toutes les stocker dans un dictionnaire. Par exemple, le dictionnaire des formes fléchies du L.A.D.L. (DELAF) occupe plus de 20 Méga-octets. La stratégie la plus utilisée est de ne stocker que les racines de tous les mots, puis de "calculer" leur flexion ensuite. Cela permet de ne pas avoir à mémoriser toutes les conjugaisons

des verbes, ni les accords en genre et en nombre des noms et adjectifs. 47 entrées sont ainsi économisées pour chaque verbe en français. L'algorithme utilisé est le suivant: le correcteur orthographique lit chaque mot du texte. Il compare ensuite ce mot avec ceux de son lexique. S'il ne trouve pas de correspondance directe, il recherche parmi les mots qui lui sont connus ceux qui sont le plus "proches". Quelques méthodes utilisées sont:

- une correction phonétique: le mot inconnu est phonétisé. Par exemple le mot "**habytation**" est transformé en "**abitassion**". Puis un certain nombre de règles sont utilisées pour transformer le mot de façon à le mettre en correspondance avec un mot connu. Cette méthode permet de corriger facilement toutes les erreurs sur les accents, les terminaisons en "tion", les lettres dédoublées (comme dans "élégamment"), les "ph", les "y" etc...
- une correction morphologique: le mot est décomposé est syllabes, puis des règles substituent une syllabe par une autre,
- une correction en recherchant la distance minimale par rapport aux mots connus; le mot fait l'objet de substitutions successives le "rapprochant" d'un mot connu,
- d'autres méthodes, qui essaient de détecter les erreurs les plus fréquentes: inversion de caractères, dédoublement de caractères...

Les correcteurs orthographiques du marché ne corrigent que les erreurs lexicales. Leur nombre limité d'entrées (quelques dizaines de milliers de mots) fait qu'ils dépassent rarement 95% de réussite.

2 Erreurs grammaticales

Les correcteurs précédents ne détectent aucune erreur dans une phrase comme:

Les belles jouent de Marie sont roses.

qui est pourtant incorrecte. Seule une analyse syntaxique de la phrase permet de corriger ce genre d'erreurs. Un analyseur syntaxique ne convient pas, car la réussite de son analyse dépend de la justesse syntaxique de la phrase. Illustrons ceci par l'exemple précédent. En simplifiant, un analyseur syntaxique cherche à analyser une phrase sous la forme

Phrase -> Sujet + Verbe + Objet

Sujet -> Groupe Nominal

Objet -> Groupe Nominal

Groupe Nominal -> Article (+ Adjectif) + Nom

Notre analyseur va commencer par rechercher un sujet, donc un groupe nominal qui est constitué d'un article, facultativement d'un adjectif et d'un nom. Il va donc commencer sont analyse de la manière suivante:

Sujet -> Les belles

Verbe -> jouent

par contre il va échouer lors de la constitution de l'objet, qui serait *de Marie sont roses*, car le verbe *jouer* n'admet pas d'objet du type *de + quelque chose* dans notre grammaire. Nous voyons clairement que l'analyseur syntaxique ne peut pas détecter l'erreur sur *jouent*. Nous nous sommes capables de la détecter en comprenant le sens de la phrase, ce qui nous permet d'isoler le sujet (*Les jouent de Marie*), le verbe (*sont*) et l'objet (*roses*). Ayant isolé chaque structure grammaticale, nous pouvons détecter l'erreur sur *jouent*.

A notre connaissance, il n'existe actuellement aucun correcteur syntaxique.

iii. Structuration des textes

Un texte d'un seul tenant est inutilisable dès qu'il dépasse une certaine taille. C'est pourquoi DARWIN permet d'organiser les textes de façon hiérarchique en documents, sous-documents etc... Cette étape est bien entendu manuelle.

iv. Indexation des textes

L'indexation est automatique. Elle peut se faire de deux manières:

- une indexation immédiate, qui se fait document par document,
- une introduction massive, qui permet d'indexer toute une suite de documents en une seule fois. Cela permet par exemple le réaliser l'indexation d'une grosse base textuelle la nuit, sans aucune intervention manuelle.

II.2.b Questionnement des bases textuelles

L'utilisateur peut questionner directement la base de données à partir d'un ou plusieurs mots ou expressions. DARWIN répond en deux temps:

- il donne une liste des expressions contenues dans la base de données et proches de la question posée. Cette liste est acceptée ou modifiée par l'utilisateur.
- Après validation de la liste des expressions acceptées, il propose une liste des textes trouvés. Il sont présentés classés par ordre de pertinence, les plus intéressants apparaissant en premier.

Les mots ou expressions peuvent être tronquées (par exemple **lingui*** permet de rechercher tous les concepts commençant par **lingui** (les réponses peuvent donc être **linguiste, linguistes, linguistique**...) . Cette possibilité permet entre autre de faire des recherches sans se préoccuper des problèmes de flexion (singulier/pluriel, masculin/singulier).

Il est de plus possible de faire des combinaisons logiques entre les mots et expressions avec les opérateurs **et** et **ou**. Les requêtes suivantes sont valides:

linguisti* ou informatiqu (pour rechercher ce qui parle de linguistique et d'informatique).

conseil régional et Bretagne (pour rechercher ce qui parle du conseil général breton).

III. Exemples de bases textuelles utilisant DARWIN

Voici quelques exemples d'applications de DARWIN:

- R.A.T.P. (Régie Autonome des Transports Parisiens): gestion du personnel (sur VAX/VMS),
- OUEST-FRANCE: plus de 1 Go de textes (120 000 articles) sur des compatibles IBM-PC en réseau,
- journaux: LIBERATION, TELEPOCHE, LE MONDE INFORMATIQUE, LA CHARENTE LIBRE...
- LA REDOUTE: tissusthèque, documentation générale sur IBM VM/CMS,
- GRÜND: système rédactionnel de dictionnaire d'artistes...
- LES MUTUELLES du MANS: réglementation des agences (1000 sites), documentation de l'infocentre sur BULL DPS 6000.
- COURTAULD, GOUVERNEMENT FEDERAL DU CANADA: version anglaise de DARWIN,
- DER SPIEGEL: version allemande de DARWIN.

IV. Bibliographie

Boons Jean-Paul, Guillet Alain, Leclère Christian, 1976. La structure des phrases simples en français: les verbes intransitifs. Droz, Genève.

Courtois Blandine, 1984, 1989. DELAS : Dictionnaire du LADL pour les mots simples du français. Rapport technique du LADL, Paris.

Gross Gaston, 1986. Typologie des noms composés, Rapport A.T.P. Nouvelles Recherches sur le langage, Paris XIII, Villetaneuse.

Gross Gaston, Jung René et Mathieu-Colas Michel, 1987. Noms composés. Rapport n°5 du Programme de Recherche Coordonnées Informatique Linguistique, CNRS, Paris.

Gross Maurice, 1968. Grammaire transformationnelle du français : 1. Syntaxe du verbe. Cantilène, Paris.

Gross Maurice, 1981. Les bases empiriques de la notion de prédicat sémantique. Langages n°63. Larousse, Paris.

Gross Maurice, 1982. Grammaire transformationnelle du français : 2. Syntaxe du nom. Cantilène, Paris.

Gross Maurice, 1986. Les adjectifs composés du français. Rapport n°3 du Programme de Recherche Coordonnées Informatique Linguistique, CNRS, Paris.

Gross Maurice 1988. Sur les phrases figées complexes du français. Langue française n°77 "Syntaxe des connecteurs". Larousse, Paris.

Gross Maurice, 1989a. La construction de dictionnaires électroniques. Annales des télécommunications, tome 44, CNET.

Gross Maurice, 1989b. The use of finite automata in the lexical representation of natural language. Electronic Dictionaries and Automata in Computational Linguistics, LITP Spring School on Theoritical Computer Science. Springer Verlag, Berlin-Heidelberg.

Gross Maurice, 1989c. Grammaire transformationnelle du français: 3. Syntaxe de l'adverbe. Cantilène, Paris.

Guillet Alain, Nahmani Stéphane, Pelletier Béatrice, 1989: Dictionnaire du LADL. Quelques classes remarquables. Rapport technique n°16 du LADL, Université Paris 7 et CORA SA.

Knuth Donald, 1973. The Art of Computer Programming, vol. 3 : Sorting and Searching. Addison-Wesley Publishing Company.

Laporte Eric, 1988. La reconnaissance des expressions figées lors de l'analyse automatique. Langage n°90: "Les expressions figées". Larousse, Paris.

Mathieu-Colas Michel, 1987, Variations graphiques de mots composés. Rapport n°4 du Programme de Recherche Coordonnées Informatique Linguistique, CNRS, Paris.

Maurel Denis, 1989. Reconnaissance de séquence de mots par automate. Thèse de doctorat en informatique, LADL, Université Paris 7.

Perrin Dominique, 1989. Automates et algorithmes sur les mots. Annales des télécommunications, tome 44, CNET.

Roche Emmanuel, 1989. Automates des phrases simples du français. Rapport de stage, CERIL, Université Paris 7, Paris.

Sabah Gérard, 1989: L'Intelligence Artificielle et le langage: répresentation des connaissances. Hermès, Paris.

Salkoff Morris, 1973, Une grammaire en chaîne du français. Dunod, Paris.

Silberztein Max, 1989. Dictionnaires électroniques et reconnaissance lexicale automatique. Thèse de doctorat en informatique, LADL, Université Paris 7.

Sommaire

I.La linguistique informatique. 2
I.1 Le domaine 2
I.1.a Les outils 2
I.1.b Les applications. 2
I.1.c Les réalisations industrielles. 3
I.2 Le problème de la compréhension 3
I.2.a.Les trois niveaux d'analyses de la langue écrite. 3
I.2.b.La compréhension. 4
I.2.c.Les principales difficultés 4
II.Les bases de données constituées de textes intégraux 6
II.1 Les systèmes à thésaurus 6
II.2 Les systèmes indexant directement les textes 7
II.2.a Constitution des bases textuelles 8
i.Introduction des textes 8
ii.Correction des textes 8
1.Erreurs lexicales 8
2.Erreurs grammaticales 9
iii.Structuration des textes 9
iv.Indexation des textes 10
II.2.b Questionnement des bases textuelles 10
III.Exemples de bases textuelles utilisant DARWIN 11
IV.Bibliographie 12

THE ECONOMIC ASPECTS OF DEVELOPING AND MARKETING FULL TEXT DATABASES

by

Mark Hepworth
Information Consultant
Financial Times Profile Information
P.O. Box 12, Sunbury on Thames
Middlesex, TW16 7AH
United Kingdom

What I would like to do today is to give you a brief overview of PROFILE Information, the process involved in getting and delivering a full text source of information online and the primarily economic issues involved. The process is broken down into collecting the data, designing and developing the database, delivering and selling the database.

PROFILE INFORMATION

PROFILE Information originated as a full text news database primarily serving the media industry. Media users include the major international television companies (BBC, ABC) and major newspapers and journalists. PROFILE now holds the full text of all the UK major national newspapers (FT, Independent, Guardian etc), newswire sources such as Associated Press, and international publications such as The Economist, Business Week, and The Washington Post.

PROFILE has since expanded it's service primarily to serve the professional sectors (Management Consultancy, Advertising, Marketing, and also the Financial Analyst and Researcher market). This has meant adding market research reports (Euromonitor, Keynotes etc), company house information on all 2.2 million limited UK companies, and company news from a range of European Publications.

I have broken down the process of providing access to a full text online database into the following sections:

1. Choosing the relevant sources;
2. Collecting the data;
3. Designing and developing the database;
4. Delivering and selling the database.

In general terms once chosen, the sources of information, which are increasingly in an electronic format, are transmitted to the online company. However, if data is not already in electronic format the hardcopy has to be scanned. This is then run through checking programmes and then through programmes that structure the data. The data is then loaded and indexed. Delivery to the customers is primarily through users dialling up via local telephone services or via International data communication networks such as Telenet. Selling the database will be through a direct sales team, distributors, such as Data Arkiv or Aftenposten or through gateways such as ESA or electronic mail services. In addition sales are generated through advertising and trade shows. All of this requires significant investment in computer power and storage, and manpower. Nexis in the States has two hundred salespeople. At PROFILE we would have one third of the staff in Sales and Marketing, one third in data preparation and one third is system development.

Choosing the Relevant Sources

Online databases tend to conform to two major types:

1. The large host, such as Dialog, carrying numerous, diverse sources, and
2. The niche database.

The first question therefore is what type of database do we want to be, are we for example amassing as many sources as possible to become the generic one stop information source, or highly focused. If one is targeted or niche oriented, how is that niche defined, what are the job functions of our target users, what information and in what form should it be delivered to facilitate these functions. Is there a base of information required, for example, of

company, market, product and industry information with a leaning towards specific vertical markets.

We are then drawn into the complex debate of the value of information. The database provider has to decide whether the cost of hosting the data can be justified by either demand by a mass market or demand by a niche market. The latter could be through a more "tailored" service enabling a higher margin.

A company will create it's own database or have a private file on a bureau of internal company news and yet use a database such as PROFILE to get public news about it's competitors and clients. The latter may then be redistributed alongside the internal corporate information via internal corporate systems or on request from the librarian. For a company to go for the internal solution, that solution needs to be tailored and satisfy a "concentrated need". However an online host may see a "concentrated need" that could be satisfied by "filleting" the database and delivering that sub set of information via an SDI service, by fax or E-mail or perhaps by producing a CD-ROM product. The latter may be facilitated by providing a vertical market or task driven interface.

The "host" type solution may imply aiming for sources that encourage as much traffic or requests as possible. However, the most profitable situation occurs where the information is of high value, storage costs are low and transactions are few. Databases have gone out of business because their computing costs increased at a greater level than their actual margin. More sources were added involving more computing costs.

Collecting the Data

The main two questions here are: is the data in an electronic format and what is the quality of the data. Increasingly publishers are using the "direct input" computerised editorial systems, such as ATEX or Norsk Data. This means that data can be transmitted to the online host in a digitalised format containing "flags" or codes that can be used in the loading programme. If the hard copy has to be scanned using optical character recognition costs are significantly higher.

The quality of the data varies in terms of the editorial quality and also the quality and structure of the data. At PROFILE, for example, we employ sub-editorial staff to write informative headlines, to restructure tabular material, to draw together duplicated news wire stories. With regard to the structure of the data, how extensive and how regular is the news editorial systems coding. Are headlines, author field, end of text markers always apparent. This in turn has bearing on how sophisticated and tailored the collection programmes have to be and also how much manual intervention and checking is required.

Designing and Developing the Database

Initial questions about the database software may be it's appropriateness for textural information versus numeric information. What proportion of the textural information is numeric, what costs would be involved in processing numeric information on a primarily textural database? Certain software may be expensive or CPU intensive to load but cheaper to access if the indexing chosen addresses the needs of the users. In general the larger the database the greater the processing costs, but depends also on the volume of transactions. Updating is also a key issue, to update one huge databases may mean rebuilding the entire index. Therefore there is an implied need to break the database up into multiple databases with multiple indexes. If, for example, you had one index and kept adding pointers to other sub indexes, the process would gradually get slower and slower and increasingly expensive.

Also involved on database design side of a full text free text search system is whether all words really are searchable. If one could search for "stop words", (and, or, of, the etc.) this would be expensive in CPU terms. It may not be useful to be able to search on common "banners" or disclaimers occurring in every item and by excluding these could save money. The degree of searchability therefore is a relation between how useful to the user are the searchable fields and how expensive does this make an individual search. Many online databases do not adequately reflect users needs in terms of searchability or manipulation of the data. One way round this is to produce focused PC interface software that will help manipulate the data on the database and achieve specific goals. This avoids having to include all these facilities on the mainframe interface.

Delivering and Selling the Database

When an online company is setting a cost for the online product significant underlying ratios are apparent. Computing costs can be as much as 30%, royalty payments to the data providers could be again 30% and cost of sale i.e. marketing, promotion and salespeople could be another 30% This leaves remarkably little profit margin and

results in a juggling act between these three factors and raises questions about how can we add value to the data and how can additional revenue be brought in from the same computing cost i.e. how can the margins be increased. For example, charged multiple use of the same information in an organisation would increase the text vendor's margin. Early investment and high costs may imply a high degree of profitability in the future.

To reduce the cost of sale the full text vendor may decide to do their business through third parties. They may be distributors or gateways through other online databases or other means of delivery such as electronic mail companies. Access to full text databases is attractive to the occasional researcher who may use electronic mail for communication purposes. The negative aspect of this approach is that the online vendor loses control over their product and does not have as close a relationship with their third party customer as they would have with a direct customer. Usage may therefore be limited but could have the reward of broad penetration of markets the online vendor would not otherwise reach. Distributors also need constant management to maintain their commitment.

With the direct customer, there is the proverbial debate about who one is selling to, the end or the intermediary. Full text was envisaged as appealing to the end user. However usage is probably still 60% by the intermediary or information professional. This is primarily due to three reasons, firstly full text has been successful and popular with the intermediary and secondly, the lack of products that relate and are well marketed to specific end users with specific job functions and needs and thirdly, the technical and psychological barriers of using an online service.

CONCLUSION

Managing full text online databases is therefore a juggling act, manipulating a variety of factors such as royalty payments, the cost of sale, revenue, and the underlying equations in the technical area (speed of access, size of storage, computer processing time, updating). Generally there is a need for market specialisation, understanding specific job functions and related needs. These needs will be satisfied by the online vendor through a variety of means. For example, providing information through various mediums such facsimile (FAX), electronic mail and CD-ROM. A variety of channels such as Corporate Networks and internal information systems will be used to deliver online information. It is also apparent that the textual and numeric online vendor will add value to the information they store through effective personal computer (PC) and mainframe interfaces helping users retrieve, manipulate and distribute the information.

BRIDGING THE COMMUNICATION GAP: THE CASE OF THE PORTUGUESE INFORMATION SYSTEM FOR INDUSTRY

Maria Joaquina Barrulas
Department of Information Studies
The University of Sheffield
Sheffield S10 U.K.
and CITI/LNETI

Zita Correia
CITI - Centro de Informação Técnica para a Indústria
LNETI - Laboratório Nacional de Engenharia e Tecnologia Industrial
Azinhaga dos Lameiros, Est. Paço Lumiar
1699 LISBOA CODEX - PORTUGAL

Abstract

The recent Portuguese economic developments are described, as well as the main plans that shape the future developments in Portugal. A brief overview of the industrial structure is provided and the national information infrastructure is analysed, with special reference to the National Library, the public libraries, the archives, special libraries and information systems, telecommunications, the information industry and available training programmes in Library and Information Science. The Programme for the Development of the Information System for Industry is described and special emphasis is given to the Post-Graduate Course for Information Intermediaries, which aimed at preparing qualified professionals to staff the information units of the System. In order to evaluate the impact of training provided through this Course on job engagement and performance of these information professionals, a survey was conducted and the results obtained are analysed . The conclusion stresses that, considering the present state of development of the Portuguese information infrastructure, a major contribution to reduce the communication gap is still to invest further in education and training of information professionals.

1 - INTRODUCTION - THE PORTUGUESE ECONOMIC DEVELOPMENT

The first and significant effort towards the industrialisation of Portugal occurs only after the second world war. Following the economic reconstruction movement of the European countries, the development of the Portuguese society from a rural and commercial economy, was then initiated. The bases of the national industrial infrastructure were established and are still the main support of the present structure.

During the 1960s and 1970s there was a progressive modernisation and openness to foreign markets. The electronic and mechanic industries, the pharmaceuticals, textiles and food industries sectors developed.

The 5 years previous to 1974, were years of rapid economic growth for Portugal. Between 1968 and 1973, the production increased 7% (average) per year.

As pointed out by the World Bank Report, " In terms of the world economy and its impact on the domestic situation, the Portuguese revolution could not have found a time more likely to complicate the adjustment and impede

future growth than April 1974. The oil price increase brought with it a substantial worsening of the country's terms of trade, while the downturn in Western Europe meant the slackening of demand for Portugal's exports, less earnings from tourism and perhaps most serious of all the levelling off of the demand for Portuguese workers" (1).

Between 1973 and 1977 the GNP per capita was virtually stagnant. The flux of returnees from ex-colonies and reduction of emigration increased the population of the country by 10%. Important markets for Portuguese products were lost (Angola, Mozambique).

In 1976 there was a slight recovery of the economy, measured by an increase of 8.6 % of the GNP. This tendency continued in the later years, but the OECD reported in 1981 that, at the beginning of the 80s, Portugal was in many aspects a developing country, with one of the lowest GNP of the OECD countries and with a production capacity largely insufficient to meet the demand (2).

In the 80s a success stabilisation Programme within an agreement with the International Monetary Fund (IMF), allowed the country to move from a situation of 13.5 % deficit in the external current account in 1982, to quasi-equilibrium in 1985.
The increase of growth permitted the reduction of unemployment from 8.6% in 1984 to 6.1 % in 1988. The inflation rate in the same period decreased 20 percentage points, and a surplus (overflow) of the Balance of Current Transactions allowed the repayment of part of the external debt.

2 - THE PLANNED FUTURE

The medium-term planning strategy for Portugal is defined in the Main Planning Options 1989/1992 (GOP) adopted by the Government (3). The overall development strategy of the GOP relies on a theoretical basis that considers that Portugal will be affected by the dominant evolution tendencies of the internationalisation of the world economy and the process of European integration.

It is also assumed that a correct and efficient utilisation of the Structural Funds of the EEC, will constitute one of the major components of the global development strategy.

The GOP, the PCEDED (Programme for the correction of the structural external deficit and unemployment), and the Programmes within the European Community Aid Framework, constitute the political guidelines for the development of Portugal in the near future.

2.1 - PCEDED - Programme for the Correction of the Structural External Deficit and Unemployment

A macroeconomic strategy for Portugal was set up in 1985 with a programme called "A strategy of controlled progress". This Programme has been successively revised and adapted (1987,1989) and is still a major framework to understand the "Planned Future" for Portugal (4).

The driving forces of this Programme are the modernisation and strengthening of the economy, as well as the reduction of the external dependency. It is considered that the structural adjustment of the Portuguese Economy and its preparation to the 1992 Single Market, requires a macroeconomic policy centred in three main areas:

i) The modernisation and increase of productivity, by means of a continuing investment effort.

ii) The reduction of inflation.

iii) The reduction of the financing of the Public Sector.

The aim is to reach 1992 with an inflation rate at approximately the average of EEC countries, with an acceptable deficit of the Balance of Current Transactions.

With the framework of the planning document PCEDED, several projects were adopted and are being implemented within the social,

cultural and economic areas of activity of the country.

2.2 - Industrial Structure

As a result of the historical and economic development process of the Portuguese society in the last 40 to 50 years, different industrial capacities related to different types of production originated an industrial structure covering various industries with technologies and companies of several generations.

The small and medium enterprises (SME) have a significant contribution to the GAV (gross added value) in the transformation industries, ca. 45%. The SME and even the very small enterprises with less than 6 employees represented in 1988, 99,2% of the total of the industrial enterprises. Only 264 industrial enterprises employed up to 500 workers and about 86 had 1000 or more. The figures for the industrial exports show also that more than 50% of the country's exports are done by SME (5).

The major difficulties faced by these enterprises and by the industrial enterprises in general can be summarised as follows:

- few innovative entrepreneurs as shown by the low levels of investment in new areas;
- few productivity, as a result of the insufficient professional training and the use of obsolete equipment and technology and also of the lack of adequate management techniques;
- weak financial structures;
- very weak links with Universities or Research Institutes;
- lack of technical, technological and market information.

2.3 - PEDIP - (Structural Programme for the Development of Portuguese Industry)

Taking into consideration the situation above described, the Portuguese authorities defined as the fundamental goal of the country's industrial policy the preparation for the 1992 European Open Market.

One of the major operational programmes within the Development Plans for Portugal is PEDIP, (6) not only because of the relatively high amounts involved, but also because of the leading role of industry in the process of modernisation and development of the country.

The objectives of this Programme, approved by the European Commission and funded by the Structural Funds of the EEC, are:

i) To revitalise the present industrial base by giving financial support to the existent enterprises which have solid economic structures;

ii) To develop new industries;

iii) To eliminate or reduce the comparative structural disadvantages by means of: Infrastructure support - creating basic technological infrastructures and launching professional training programmes; Financial support and support to dynamic factors of competition like, improvement of quality, industrial design, productivity missions.

Several funding schemes have already been implemented and a great deal of improvement is expected from this major Programme, during its 4 years duration, until 1992.

3 - THE NATIONAL INFORMATION INFRASTRUCTURE

The definition of the Information Infrastructure concept involves some difficulty. For GRAY (7), infrastructure is a term to avoid when National Information Policy is concerned, because it means different things to different people: " Some equate it with the whole organisational structure of information services, others with general services as distinct from the specialised services that are provided within areas of economic, social and cultural development".

ZURKOWSKY (8) and the American Information Industry Association, use the neologism "Infostructure" to encompass all the elements necessary to support the information handling capacities of the country. The objective of the infostruture being "to create, communicate and deliver information useful to all economic, social and political activities of the country".

For the purpose of our study we consider the Information Infrastructure as including the following basic components:

- Information resources
- Organisations
- Information Industry

i) By information resources is understood:

- the information sources of all kinds, recorded knowledge in every physical support;

- the human specialists devoted to information-related activities;

- the technologies used as a means of recording, delivering and communicating information;

- the telecommunications devices that constitute the channel for the delivery of information .

ii) As organisations of the information infrastructure we consider:

- the institutions traditionally devoted to the activity of collecting, handling and making available the recorded knowledge: Libraries, Information and Documentation Centres and Archives.

- the more recent organisational environments created with the main purpose to enhance the knowledge diffusion and utilisation: Science Parks, Innovation Centres, Technological Centres.

iii) The information industry in this context is restricted to information technology manufacturing and information provision activities, excluding the mass media systems.

In order to draw an up-to-date picture of the present Portuguese Information Infrastructure, a survey was done as part of the work that is being carried out by one of the authors, at the University of Sheffield. The major findings are reported below.

3.1 - National Library

A Programme to reorganise the National Library started in 1985. As a result of the studies undertaken, a decision to create a National Bibliographic Database -PORBASE- was approved by the Government. The GEAC system was chosen and the process to install the equipment started in 1987. The PORBASE is now available online having around 200000 records, corresponding mainly to the National Bibliography (1950-1990). A process of retrospective conversion is in due course. It includes also records from the holdings of other Libraries all over the country.

Holding ca.two million volumes, the National Library has developed its collection specially in the fields of Arts and Humanities while it has a rather small collection of Science and Technology. With a small budget, two thirds of which goes directly to pay salaries, the NL has no acquisitions policy. The collection is built mainly through the Legal Deposit (which increases the collection by a proportion of 10 to 12 thousand volumes per year). The lack of trained human resources is considered by the management the major problem of the National Library. In a total of around 300 staff, only 33 are qualified Librarians and at present no computer specialist is employed by the Library, despite the equipment installed (a mainframe computer, several micros and peripherals). This shortage of qualified staff prevents the library from opening for example, the Audio-visual section. There are also no research staff, albeit the NL participates in some international cooperative projects,

namely in EEC Action Programme for Libraries.

3.2 - Public Libraries

There is no tradition of regular use of Public Libraries in Portugal. The first time that the public library designation appeared was in the name of the Real Biblioteca Pública da Corte, created in 1796. Several other public libraries were created in the 19th century, but mainly because the extinction of the religious orders in 1834 resulted in a need for the preservation of the bibliographic holdings of the monasteries and colleges. With the advent of the Republic, in 1910, Libraries started to be seen as "instruments to destroy ignorance" . A law published in 1911 envisaged the creation of "Popular Libraries" in every municipality. In 1919 a survey showed that there were libraries in 68 municipalities, 12 of which were still in an organisation phase and 37 having less than 2000 volumes. In 1958 another survey showed that from the 273 municipalities in the country, only 66 had a functioning library (ca. 25%), while in 1982 (almost a quarter of a century later) this proportion has increased only 10% : from the total of 275 municipalities, 97 have a library.

In general the municipalities with libraries are those of the more developed areas of the country and the common feature of these libraries is: small and/or old collections, obsolete equipment and no qualified staff.

Considering this scenario, the Secretary of State of Culture set up in 1986, a Task Force with the objective of preparing the basis for a "coordinated intervention in the Public Libraries" (9).

Following the measures proposed by the Task Force, a programme to establish a Public Libraries Network is now in operation. This Programme is based on a cooperative action between the Local Authority (at municipality level) and the Instituto Portugués do Livro e da Leitura - IPLL (the national coordinator body) and is aimed to cover one third of all the municipalities during its five years duration. It is assumed that by that time "a point of no return" will be achieved and the imitation effect will reach the rest of the country. Within the Programme, full support is offered to the Local Authorities that declare an interest in creating or improving their public library, provided that some basic conditions are accepted:

- 50% of the total costs involved must be supported by the Local Authorities;

- the library must follow the technical specifications decided by IPLL, concerning spaces, technical procedures (e.g. lending services and open access), and employ qualified staff.

3.3 - Archives

The preservation of archive material is of considerable importance not only for historical research purposes, but also as an instrument of development and progress. The inter connection between the resources of the past and the concrete situations of the present is a factor to take into consideration when planning the future.

One of the actions included in the GOP document mentioned in section 2. intended to contribute to the "Promotion of the National Identity Values", is the development of the National Archives Network.

The National Institute of Archives was created in 1988 having as its mission the establishment of a network of all archives collections, on a district basis.

The first step to build the Network is the implementation of standardised procedures. A software application using the UNESCO CDS-ISIS system was tailored specifically for the Network and is in the implementation phase. This will help to follow the same methods of coding and storing the materials by all the archives. Microcomputers were supplied and basic computer training provided. Again, the lack of trained human resources is the major constraint.

3.4 - Special Libraries and Information Systems

The special Libraries and Information / Documentation centres existing in the public and private organisations are in general the more advanced information systems in the country. The development of information services specially tailored to meet specific information needs of particular user groups, occurred first in the Scientific and Technical Information area. Although other information services have been developed in recent times, the STI field is still the one with more national coverage.

In the Ministry of Industry and Energy (MIE), several specialised organisations set up their own systems and started creating in-house databases.

In general, the information centres are created to meet the needs of the organisation's internal users, but later develop to the more extensive purposes of providing information also to external users. This was the case of the National Laboratory of Engineering and Industrial Technology (LNETI). The Centre for Scientific and Technical Information for Industry (CITI) of LNETI, whose mission is mainly to satisfy the research and technical staff's information needs, developed in 1985 the first bibliographic database in the field of industry and energy available online in Portugal. This database, is the online catalogue of the CITI libraries and aims to cover also all the literature produced by Portuguese researchers in the field of industry and energy and is available outside LNETI.

We refer the example of LNETI, but we know that more or less developed and technologically equipped Information Centres exist in the various Ministries, but mainly concentrated in the Central Administration Departments.

3.5 - Telecommunications

In the past 5 years, a great effort to modernise the telecommunications system and to introduce new technologies was done in Portugal.

The digital network system has been gradually implemented and by the end of 1993 is expected to cover all the country, in substitution of the analogic system. A project to implement the ISDN is planned in three phases:

- the first phase called a pre-ISDN, from 1989-1990, offering services like: videoconference, telephone with sophisticated facilities, high speed access to VIDEOTEXT and MHS networks, (Message Handling Systems) as well as virtual networks to private systems.

- the second phase, from 1990-1991, will provide 1st generation ISDN services, like: highly sophisticated telephone, digital access to Telepac(PPSN), Videotext and MHS, intercommunication between workstations and PC's, data communications at 64Kb/s.

- the third phase, from 1992 onwards, will provide 2nd generation ISDN services like: videotelephone, videoconference, digital portable telephone, access to all services at 64Kb/s (10).

The Public Packaged Switched Network (TELEPAC) was commercially available in 1985 with 3 switch nodes in Lisbon, Porto and Coimbra, and by the end of 1989, 27 nodes were installed covering the continent and the adjacent islands (Madeira and Azores).

According to international indicators the level of utilisation of data telecommunication in Portugal is still relatively low, however it triplicated from 1985 to 1987 and the annual growth rate is greater than the average of EEC countries.

The Videotext system and POS (Point of Sales) and ATM (Automated Teller Machines) services are the main causes of the increase of volume in data transfer. The Videotext system started in 1989 and was primarily oriented to the professional market. In February 1990, the use of the system was measured in 4.604 hours connection time. The

forecast for the near future is a rapid growth, mainly because of the penetration in the residential market (11).

3.6 - Information Industry

A recent report elaborated for the EEC, within the IMPACT Programme (12), refers that the information industry and information services market in Portugal are still in an incipient phase of development. The information related activities were, until recently, identified only with Libraries and STI Centres. Several initiatives from other areas are enlarging the borders of the sector, but a lack of identity of the different players in the field, who do not see themselves as being involved in the same business, is one of the reasons of the reduced size of the information industry in the country. However, it is reiterated that a great potential exists in some areas.

The survey shows that Tourism, Environment, Services to Local Communities, Geographic Information Systems are the areas where a major contribution to the development of the information market in the short term, is expected. This conclusion took into consideration the evaluation of the projects already in due course, the resources located to them, as well as the potential consumers market.

The report also emphasises that the lack of qualified human resources is a major constraint to the development of the information industry and information services market in Portugal.

3.7 - Training Programmes

The officially recognised training programmes in Information were conceived as specialisations to be built on top of other academic qualifications rather than a scientific ground on its own. Therefore, no graduate program is available on this area. Training programmes are to be found providing professional orientated courses on the top of secondary education, or post- graduation courses open to every graduation background.

Professionally oriented courses are taught in some Secondary Schools in Coimbra, Lisboa and Porto within the area of Humanities and convey specialised training concerning Information Technology, Sociology of Information, Management and Information Techniques. Two branches are available, one for preparing Library professionals and other for preparing Archive professionals, both at an assistant level. Before the recent integration of these curricula into the National Educational System (1989), these courses were promoted and supported by BAD - Associação Portuguesa de Bibliotecários, Arquivistas e Documentalistas, one of the Portuguese professional Associations.

The Post-Graduation Courses in Librarianship were created in 1983 in order to replace the former post-graduation course that had a run since 1935 with a rather conservative curriculum which aimed mainly at the training of archive specialists. These two-year courses are run by the Universities of Coimbra, Lisboa and Porto, the first and second semesters consisting of a joint curriculum both for librarians and archivists, while the third and fourth semesters are taught separately specialised curricula for the two different branches. The core disciplines are Management and Informatics (Introduction to Computer Science) for both branches, Subject Indexing and Cataloguing (for Librarians) and Archivology and Paleography (for Archivists). A rather strict numerus clausus does not allow the enrolment of more than twenty students for the librarians option and ten for the archivists, annually.

4 - THE PROGRAMME FOR THE DEVELOPMENT OF THE INFORMATION SYSTEM FOR INDUSTRY

In the previous paragraphs the main points of the movement towards the modernisation of Portuguese society, were outlined. A chronological picture of the economic development experienced in the last four or five decades, and a general overview of the present information infrastructure was given.

The Programme for the Development of the Information System for Industry undertaken in 1987, under the auspices of the Ministry of Industry and Energy (MIE), must be analysed in the context of the above framework. This experimental Programme has already been described in detail elsewhere (13,14). Briefly, the Programme aimed to create Information Nodes in several Industrial Associations -IA (the Portuguese equivalent to Chambers of Commerce). The Information Nodes should be staffed by two qualified Information Intermediaries and furnished with a Microcomputer linked to the PPSN, fax and telex facilities and with a basic reference collection. Financial and technical support to be given by CEDINTEC - Centre for Technological Development and Innovation (an agency for the encouragement of technical innovation of the industry), and by LNETI/CITI, respectively.

An Information System is essential to the overall project to modernise the industry, and it is always mentioned as a priority in policy documents. In the MIE, several other projects to develop such system were proposed but, for various reasons failed.

We think that the uniqueness of this Programme depends on the three prerequisites considered the key factors to success:

- to create a light and flexible structure, with the technological facilities to allow the access to world-wide information sources;

- to involve the potential users in the project;

- to rely heavily upon the ability of the Information Intermediary to bridge the communication gap between the information sources and the end user.

The participation of the IA proved to be a useful means of bringing the system closer to users. And the configuration designed for the Information Nodes seems to have been the more suitable, if we consider the funds available for a three years Programme.

The Information Intermediary was seen as key player in the system. Considering the Portuguese Industrial structure, the system was intended to meet the information needs of small and medium enterprises, which were described above (see 2.2), in the different regions of the country.

To promote effectively the transfer of information to this category of users, requires specific qualifications such as the ability to:

- seek, identify, select, process and present information in a form adapted to the specific needs of users in industry:

- guarantee the efficient transfer of information from all sources available in the country and abroad;

- identify/survey the information needs resulting from the activity of industrial firms;

- contribute towards the identification of existing information resources at national level which can satisfy the information needs detected.

Facing the lack of training opportunities for information specialists, as seen above (see 3.7), it was decided to organise a training programme specifically aimed to prepare the Intermediaries who would staff the Nodes of the System.

5- THE POST-GRADUATION COURSE FOR INFORMATION INTERMEDIARIES

The training programme developed by LNETI/CITI in collaboration with the Department of Information Studies of The University of Sheffield has already been analysed in various occasions (15).

Because the impact of the course is far beyond the initial objectives and for reasons of clarity, the main lines of the course are described again.

The course is structured in two terms with a total of 750 teaching hours, theoretical,

practical and tutorial. The broad areas of the curriculum are:

i) Information in its economic, political, social and cultural context;

ii Information processing and application of information technology in developing and implementing information systems;

iii) Information resources;

iv) Information transfer and technology transfer.

A very wide range of topics was covered by the 13 modules of the course as it is outlined in Table 1.

Table 1

Modules of Term 1	Modules of Term 2
Intensive course in English	Applic. of Microcomp. in Info Management
Information Science	Business Information
Information Marketing	Info. Management
Information Processing and Microcomp. Applications	National Info. Resources
Information Resources	Research Methods
Technology Transfer	Info. Needs and design of Info. Services
Expert Systems	

An evaluation was carried out after the first course, held between October 1987 and April 1988, and a number of changes to the programme were recommended (16). Some of the suggestions were implemented in the next two courses: for instance the presence in the first month of the course, of representatives of the information providers and information users from business and industry to discuss the problems of information provision and use. Also the introduction of a new module concerning the business scene in Portugal including the treatment of: how Portuguese business and industry operates in general terms; the different types of enterprises and the need for change towards 1992 and the open market; the main lines of economic and social policy as they affect enterprises.

From the point of view of the course organisers, the above mentioned aspects relating to the business scene and the field work done as part of the National Information Resources, are of considerable value for the future professional accomplishment of the Intermediaries. The same applies to the training given in the communication aspects involved in the information transfer process.

6- IMPACT OF TRAINING ON JOB ENGAGEMENT AND PERFORMANCE

6.1 - The Questionnaire

In order to evaluate the actual impact of the training provided through the Post-Graduation Course for Information Intermediaries on job engagement and performance a questionnaire

was prepared. It contained 5 sections with a total of 21 questions, covering items such as the actual professional situation, mode of recruitment, main activities performed, utilisation of technology and perceived contribution of specific skills acquired or developed throughout the course, to the professional performance of the former students.

The questionnaire was sent by mail to all the 35 students who had attended the first and second courses in the academic years of 1987-1988 and 1988-1989.

Although the main body of questions were closed questions, those concerning the competencies (understood as a body of knowledge, skills and attitudes) required to the fulfilment of the job and those concerning the main activities performed were open-ended, in order to provide as much information as possible on those items.

The questionnaire was first tested with the collaboration of two former students. 21 responses were received, corresponding to 60% of the total number of questionnaires issued in March 1990.

6.2 - Analysis of the Responses

6.2.1 - Professional Situation

Only one of the respondents is not working in the field due to military service obligations. The other 20 are distributed in a fairly balanced way over the private and public sectors and earning salaries from Esc. 70.000$00 to 300.000$00 (see table 2). Both private and public sector jobs are based heavily on the services sector. 40% of the respondents are on their second job after having finished the course, and 35% of this group are performing the job of Information Intermediaries.

Table 2: Job Sector and Wages

Wages (Portuguese Escudos)	publ. %	priv. %	coop %
200.000/299.000		10	-
150.000/199.000	5	10	-
100.000/149.000	35	20	-
70.000/ 99.000	5	10	5

6.2.2 -Recruitment

The mode of recruitment used was mainly direct invitation, 40% (see Table 3) and interview was the most common method of selection, 70% . When asked if they had mentioned their post-graduation course during the selection process, 90% answered affirmatively and among these, 70% think that it was decisive for their final recruitment.

Table 3: Recruitment

Advert	30%
Invitation	40%
By indication of	30%

6.2.3 - Competencies Required

70% of the respondents mention as competencies required by employers for the fulfilment of the job, a graduation plus additional specific skills and/or attitudes. 30% mention a graduation plus specifically the Post-Graduation Course for Information Intermediaries without further requirements. This leads us to believe that the course was regarded by employers as containing the full range of knowledge, skills and attitudes required for the job they were offering. Table 5 shows respectively the skills and attitudes required by employers as a supplement to the graduation: computer skills, and the interpersonal communication skills are the most popular requirements.

Table 5: Competencies

Skills	%	Attitudes	%
Computing	45	Ability to adapt	10
Forg.Lang	10	Ability to Commun	25
Telecommun	5	Ability to Innovate	10
		Leadership	5
		Team spirit	10

6.2.4 - Main Activities Performed

Among the activities listed in Table 6 (several answers were allowed), retrieving information for one's own utilisation is mentioned by 90% of the respondents and participation in meetings and report writing is referred by 80%; retrieving information to provide to someone else and contacting people outside the organisation comes next, 70%. However, when asked which activities were most frequently performed, retrieving information to provide to someone else comes first, followed by answering information requests. The remaining activities mentioned by the respondents are highly scattered according to specific requirements of the different organisations, where the activities are performed.

Table 6: Activities performed

Contacting people inside the Org.	55%
Contacting people outside the Org	75%
Answering Info. requests	70%
Retrieving Info. to users	75%
Retrieving Info. for one's utilisation	90%
Index.,Catalog.,Class., Abstract.	10%
Storing Information	55%
Disseminating Info.	65%
Meetings	80%
Writing reports	80%

6.2.5 - Utilisation of Technology

Table 7 shows the level of utilisation of microcomputer packages in the activities of the Intermediaries, with high rates concerning DBMS and word-processors (80%) and spreadsheets (70%). Table 10 displays the means used daily to communicate inside and outside the organisation, with a heavy stress for personal communication inside the organisation (90%) and for telephone inside and outside the organisation (60% and 65%). Fax machines may be considered to have a relatively light use (25%) and electronic mail is little used as a means of communication both inside and outside the organisation (5% and 10%).

Table 7: Microcomputer packages

Word-processor	80%
DBMS	80%
Spreadsheet	70%
Graphics	15%
Communications	30%

Table 8: Means of communication used inside/outside the Organisation

	IN	OUT
Personal contact	90%	20%
Telephone	60%	65%
Mail	20%	20%
Telex	-	-
Fax	25%	25%
E-Mail	5%	10%

6.2.6 - Perceived contribution of specific skills acquired throughout the course

As for the answers to whether and how much the course as a whole has contributed to their professional performance, 25% state that the course is extremely useful and 30% state that it is very useful, (totalling 55%), while 5% think it is useless for the job they are now performing.

A detailed analysis of the contribution made by specific skills acquired or developed during the course (see Table 9) shows that in fact 9 out of 11 of the skills listed, collected from 50% to 80% agreement that they are extremely or very useful for their actual performance. The skills most valued are in fact the technological skills, followed by the knowledge of information sources.

Table 9: Perceived contribution of specific skills acquired through the course

Skills	Level of contrib 1	2	3	4	5	4 + 5	
Technology	-	3	-	6	11	17	85%
Info. Sources	-	3	1	7	9	16	80%
Online searching	2	3	5	3	7	10	50%
Info.for Industry	2	2	3	5	8	13	65%
Portuguese Indust	1	3	6	5	5	10	20%
Organis.of Info	-	4	4	7	5	12	60%
Present. of Info	-	1	5	7	7	14	70%
Info. Needs	-	3	2	7	8	15	75%
English Lang	-	3	3	8	6	14	70%
Communication Sk.	-	2	4	4	10	14	70%
Technology Transf	2	3	7	7	1	8	40%

7 - CONCLUSIONS

The outline made in sections 1 and 2 provides an overall view about the level of economic development in Portugal and its industrial structure, while section 3 analyses the Portuguese Information Infrastructure. They are the basis to understand the Programme for the Development of the Information System for Industry and the Post-Graduation Course for Information Intermediaries. This background, together with the results obtained from the survey to the former students, allow us to formulate the following conclusions:

i) The information infrastructure in Portugal suffers from a severe shortage of qualified information professionals. This is confirmed by the high rate of job mobility among the recently graduated information intermediaries who look for better jobs and wages and succeed in getting them.

ii) The Post-Graduation Course for Information Intermediaries has contributed both to increase the number of qualified information professionals prepared annually and to enlarge the range of information competencies available at a national level due to the innovative features of its curriculum. In general, the course is seen by the former students as an important aditional specialisation that enables them to perform information related activities in any professional environment with a high level of achievement.

iii) As would be expected from an incipient information industry and information services market, the use of information technology is also incipient, even though technological skills are highly appreciated both by employers and employees.

iv) Interpersonal communication skills are also highly appreciated and wanted by employers and prove to be one of the more frequent means of information transfer.

Considering the present state of development of the Portuguese information infrastructure we think that a major contribution to reduce the communication gap is still to invest further in education and professional training. In fact, the technology is becoming available and gradually used throughout the country, the information industry is giving its first steps but, as was widely recognised by all the key players in the Information Scene in Portugal the lack of human specialisation is the main constraint to the development of the sector in particular and the country in general.

However, the experience of the Programme for the Development of the Information System for Industry and in particular the Post-Graduation Course for Information Intermediaries show that the development of a light and flexible structure integrating technologies and qualified human resources is probably the most effective way at present, to allow the Portuguese industrial firms to access the information available world-wide in the same conditions of their foreign competitors.

8 - REFERENCES

1 - PORTUGAL: Current and prospective economic Trends.- A World Bank Country Study.- New York: The World Bank, 1978.

2 - Études Economiques de l'OCDE: PORTUGAL, 1981.- Paris:OCDE, 1981.

3 - PORTUGAL 1992: GPO's 89/92.- Lisboa:MPTA, Sec. Estado Planeamento e Desenvolvimento Regional, 1988.

4 - Estratégia .de Progresso Controlado: PCEDED, Programa de Correcção Estrutural do Défice Externo e do Desemprego.- Lisboa: Ministério das Finanças,1989.

5 - AMARAL, L. Mira. - O papel do IAPMEI no apoio às PME's no quadro da Política Industrial Portuguesa.- Lisboa:MIE, 1988.

6 - PEDIP: Objectivos, estrutura e enquadramento no regulamento comunitário e na política industrial portuguesa.- Lisboa: MIE, 1989.

7 - GRAY, John. National Information Policies: Problems and Progress.- London: Mansell Publishing Ltd, 1988.

8 - ZURKOWSKI, Paul G.- Integrating America's Infostructure. "Journal of The American Society of Information Science", 35 (3), 1984, p. 170-178.

9 - MOURA, Maria José. - Leitura Pública: Rede de Bibliotecas Municipais. Relatório apresentado à SEC.- Lisboa: SEC, 1986.

10 - BAU, Graça, FERREIRA, Godinho. - As telecomunicações CTT respondem ao desafio do mercado. Comunicação apresentada no Congresso das Telecomunicações, Porto 1989.

11- BRITO, Carlos.- Oferta pública de serviços pelo Transdata. Comunicação apresentada no Congresso das Telecomunicações, Porto 1989.

12 - CHALLENGE, Inovação e Tecnologia. - Mercado Português de Serviços de Informação. (Relatório apresentado à DGXIII-B CCE, 1990).

13 - CORREIA, Ana Maria Ramalho, WILSON, T.D. - The Information Intermediary in the Information System for Industry in Portugal. In "Information, Knowledge, evolution: proceedings of the 44th FID Congress held in Helsinki, Finland, 1988".- Amersterdam: North-Holland, 1988, 341-349.

14 - BARRULAS, Maria Joaquina, CORREIA, Ana Maria Ramalho, WILSON,T.D. - Information Intermediaries for Industry in Portugal: A training Programme and its impact. "Education for Information", 7, 1989.

15 - CORREIA, Ana Maria Ramalho. - Training Information Intermediaries in Portugal. "Outlook on Research Libraries", 10 (9), Sep. 1988.

16 - WILSON, T.D. - Curso de Intermediarios de Informação para a Indústria: evaluation report.- Lisboa:LNETI/CITI, 1988.

POSSIBLE RECOMMENDATIONS TO MINISTRIES AND OTHER AUTHORITIES BASED ON THE FOREGOING PAPERS

by

OLEG LAVROFF
5 rue Anna Jacquin
92100 Boulogne
France
(formerly of Aerospatiale, France)

1. INTRODUCTION

This paper was prepared by the Theme Coordinator after he had read the other contributions. Sections 2 and 4 were written in French and Section 3 in English. The text presented here consists of English and French versions of the complete text, each including some translations from the other.

2. AN INFORMATION NETWORK

The purpose of this paper is to draw out the many conclusions arising from the other papers presented at the meeting and, above all, to identify the tasks ahead for information centre managers and recommendations that they could make to their superiors or to Government departments with responsibility for information policy, at least in those countries where there is a national information policy. Where there is no such policy, concerted action should be taken to persuade a ministry or other official body to define one, based on the recommendations arising from this meeting.

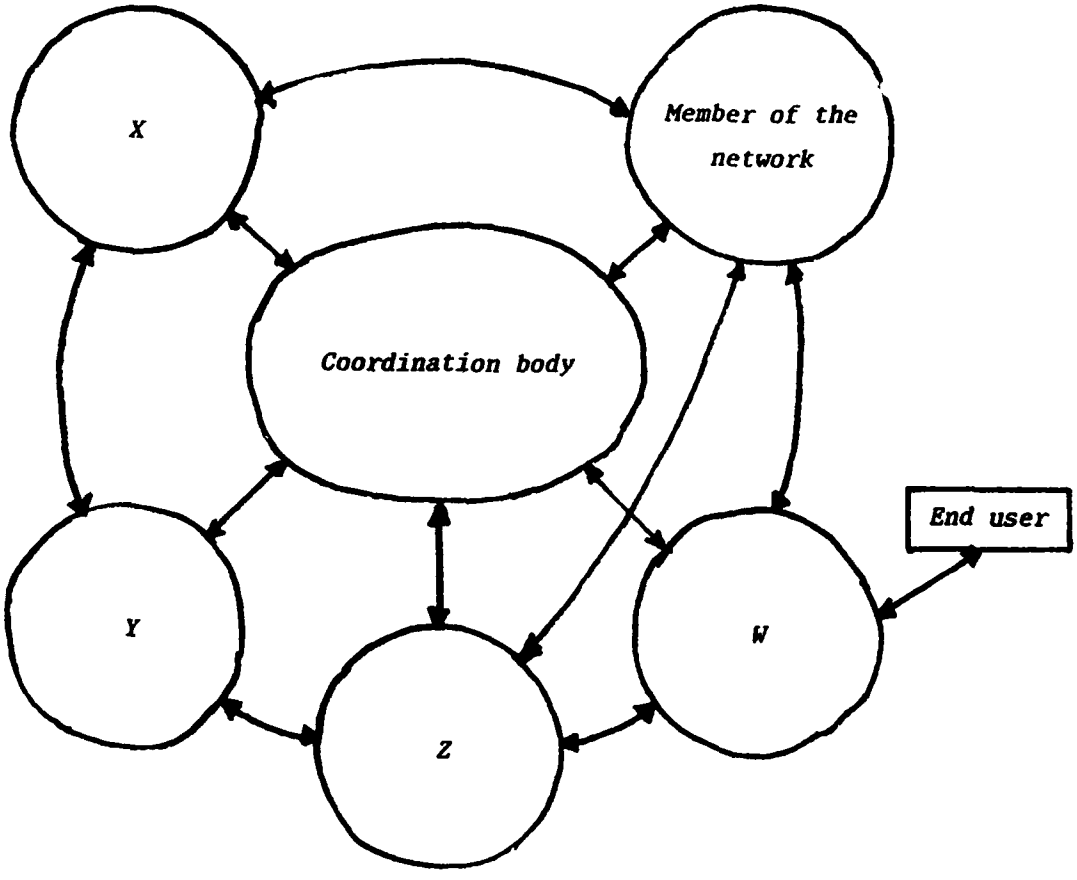
It is of course a very ambitious project, but if all information centre managers wish to be able to respond fully to the desires and needs of their end-users (who attach more and more importance to relevance of information and particularly to speedy delivery of it in their own language) it will be essential to develop and provide for their use all the necessary means (logistic, personnel, financial and training), to enable them to exploit fully and effectively the new technologies which will soon be used in research and for information transfer.

It is obvious that not all the recommendations can be applied in all countries, or even in all information centres, but they should be considered as guidelines for the definition of information policies in each country and in each information centre. We live now in a world which has a permanent need for information. This means that we should all use compatible methods and means for the transfer of information. Only if we do so, can we bridge the existing communication gaps.

To do so, we must first of all have in mind, or be able to define, guidelines for the organisation of an information

network, in which each participant should be considered both as a supplier and a user of information. Such a network can operate only on the principle of reciprocity in which the members of the network all agree to share the tasks. In this way, each member can have quick access to a mass of information at low cost. But that is possible only if the operations of the members of the network are fully compatible with one another.

SCHEMATIC REPRESENTATION OF AN INFORMATION NETWORK



This type of network can be adopted either at a national level or within an organisation or a company. It is a multiple communication network which must be totally transparent to the end-user.

Based on this plan, and on the papers presented during this meeting, the main recommendations can be classified under the following headings, each of which is elaborated below:

1. Logistic support
2. Standardisation
3. Artificial Intelligence
4. Assistance to information suppliers/users
5. Knowledge base

3. RECOMMENDATIONS

3.1 Logistic Support

Organisational

Creating a coordinating body responsible for the definition of a documentation policy and its implementation;

Defining and organising an information network.

Communication Networks

developing and introducing ISDN networks;

developing and introducing the use of optic fibres;

planning different types of communication networks: centralised, decentralised and with multiple connections.

Computer Systems

Giving priority to a high-performance and user-friendly electronic storage system connected to the existing computer application environment (in most cases, PC terminals can be connected to large host systems).

3.2. Standardisation

Using international standards for the acquisition of texts and pictures in computer systems (from the writer to the editor of such information).

Harmonising the procedure to access to various information service centers in consultation with the system designers and users.

3.3 Artificial Intelligence

Studying and developing automatic systems for indexing or abstracts (concept extraction).

Studying and developing connection tools to improve access to service centres and intelligent, multilingual and user-friendly databases.

3.4 Assistance to Information Suppliers/Users

Developing computer assisted writing systems (to deal with problems such as badly-written texts, ambiguous sentences, spell checking, etc..).

Creating multilingual terminology databanks whose applicability is to be checked by user countries.

Developing Computer Assisted Translation (CAT) systems.

Improving CAT workbenches.

Training the potential users of new technologies.

3.5 Knowledge Base

Organising and structuring the selection and acquisition of past, current and future information to be recorded in a national databank (know-how/grey matter of a country), while keeping the confidential characteristics of circulated information.

Defining the structure and contents of the databases.

4. CONCLUSION

In outline those are the main recommendations that we can make to the competent authorities. It is difficult to specify an order of priority, but if we want to bridge the communication gap in all its forms, these recommendations must be developed and used very quickly. However, when we examine the list it is clear that some actions have already been started in all the areas mentioned, with proposals for funding which should, in most cases, allow the completion and marketing of all the projects under consideration.

Why, then, should we make any recommendations?

1. Because users should work with system designers, in order to make their needs known.

2. Because all the systems developed should be usable within the existing office automation environment, should be compatible among themselves and should be able to form part of a general information network. Hence the importance of coordination by an official organisation.

3. In order to convice our authorities of the need for an official coordinating organisation charged with defining an information policy and initiating it.

Very briefly, then, these are the reasons which lead us to make recommendations that can be addressed to the appropriate authorities and to take actions, particularly to ensure the training of the personnel who will have to use these new technologies.

PROJETS DE RECOMMANDATIONS A FAIRE AUPRES DES MINISTERES ET AUPRES D'AUTRES AUTORITES SUR LA BASE DES COMMUNICATIONS PRESENTEES LORS DE LA CONFERENCE

par

OLEG LAVROFF
5 rue Anna Jacquin
92100 Boulogne
France
(Chef, retraité, de la Section Information d'Aérospatiale, Suresnes, France)

1. INTRODUCTION

Le coordonnateur du théme de la conférence a rédigé cette note après avoir lu les interventions des autres conférenciers. Les sections 2 et 4 ont été rédigées en français, et la section 3 en anglais.

2. UN RESEAU D'INFORMATIONS

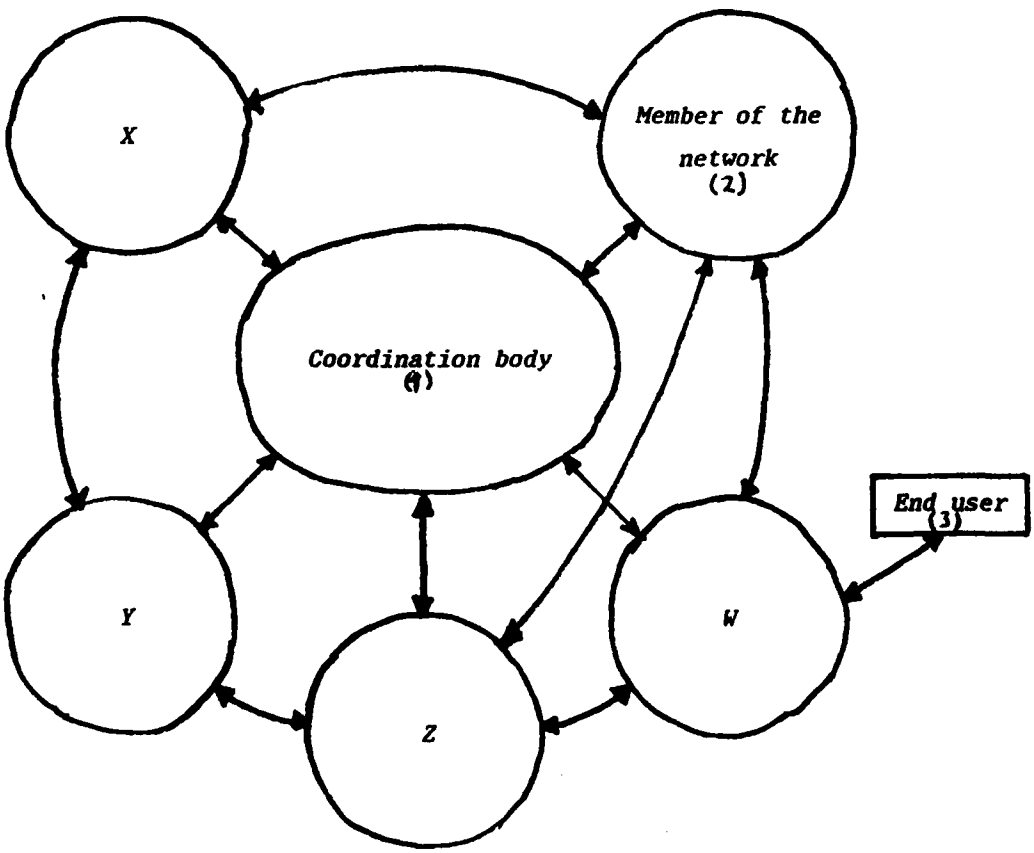
L'objet de cette intervention est de dégager les principales conclusions qui découlent de toutes les interventions faites à ce Congrès, et surtout, de mettre en évidence les efforts que devront mener les responsables des Centres d'Informations, et les recommandations qu'ils pourraient faire auprès de leurs instances hiérarchiques ou des Ministères chargés de définir une politique documentaire, dans le cas bien sur, où une politique documentaire nationale existe dans un pays. Si ce n'est pas le cas, il faut tenter par des actions concertées, de faire définir par un Ministère ou un organisme officiel, une politique documentaire nationale, basée sur les recommandations issues de ce Congrès.

Très certainement, il s'agit là d'un projet ambitieux, mais si tous les responsables des Centres d'Informations souhaitent pouvoir répondre favorablement aux désirs et aux besoins des utilisateurs finaux (lesquels attachent de plus en plus d'importance à la pertinence, et surtout, à la réduction des délais dans la fourniture de l'information dans leur langue maternelle), il s'avère indispensable que tous les moyens (logistique, humain, financier, de formation) soient développés et mis à leur disposition, afin de leur permettre d'utiliser pleinement et à meilleur profit, les nouvelles technologies qui seront utilisées à court terme dans la recherche et le transfert de l'information.

Il est évident que toutes les recommandations ne peuvent être retenues et appliquées dans tous les pays, ou simplement dans tous les Centres d'Informations, mais ces recommandations doivent être considérées comme un axe directeur dans la définition des politiques documentaires de chaque pays ou de chaque Centre d'Informations. Nous vivons maintenant dans un monde qui a besoin d'information en permanence, et cela implique des méthodologies et des moyens harmonisés reconnus et utilisés par chacun d'entre nous. Ce n'est que dans ces conditions que nous pourrons combler les lacunes dans le domaine de la communication.

Pour cela, il faut tout d'abord avoir à l'esprit, ou définir, un schéma directeur de fonctionnement d'un réseau d'informations, où chaque membre participant doit être considéré comme un fournisseur et utilisateur d'informations. Ces réseaux ne peuvent fonctionner que sur le principe de la réciprocité, c'est à dire qu'après une concertation, les taches soient réparties entre les membres du réseau. Ainsi, chaque membre pourra accéder à une masse d'informations à moindre coût et des délais plus courts. Mais cela ne peut fonctionner que dans un réseau où les moyens mis en oeuvre et utilisés par chacun soient totalement compatibles.

SCHEMA D'UN RESEAU D'INFORMATIONS



1. organisme coordonnateur *2. membre du réseau* *3. utilisateur final*

Ce réseau peut s'adapter à un niveau national ou à une organisation interne d'un organisme ou d'une Société. Il s'agit d'un réseau à communication multiple et qui doit être totalement transparent pour l'utilisateur final.

En se basant sur ce schéma de principe et sur les exposés présentés au cours de ce Congrès, les principales recommandations peuvent se répartir sur les 5 points suivants :

1) *Moyens logistiques*
2) *Standardisation*
3) *Intelligence artificielle*
4) *Aide aux utilisateurs / fournisseurs d'informations*
5) *Base de connaissances*

3. RECOMMANDATIONS

3.1 Appui logistique

organisationnel

création d'un organe de coordination responsable de la définition d'une politique de l'information et de sa mise en application.

définition et organisation d'un réseau d'information

réseaux de télécommunications

réalisation et mise en oeuvre de réseaux RNIS

réalisaton et mise en oeuvre de la solution fibres optiques

préparation de différents types de réseaux de télécommunications :
centralisés, décentralisés et à connexions multiples.

systèmes informatiques

donner la priorité à un système d'archivage electronique de haute performance et convivial connecté à l'environnement informatique existant avec ces applications (dans la plupart des cas des micro-ordinateurs peuvent être utilisés comme terminaux reliés aux grands serveurs).

3.2 Normalisation

utilisation des normes internationales pour l'acquisition de textes et figures dans les systemès informatiques (du rédacteur à l'éditeur de l'information).

harmonisation de la procédure d'accès aux divers centres et services d'information, en consultation avec les concepteurs de systemès et les utilisateurs.

3.3 Intelligence artificielle

étude et réalisation de systemès automatiques d'indexation ou d'analyse (extraction de concepts).

étude et réalisation d'outils de connexion pour améliorer l'accès aux centres et services et à leurs bases de données par des moyens intelligents, multilingues et conviviaux.

3.4 Assistance aux fournisseurs d'information et à leurs utilisateurs

réaliser des systemès de redaction de texte assistés par ordinateur (pour traiter les problemès tels que textes mal rédigés, phrases ambigües, contrôle orthographique, etc).

créer des banques de données de terminologie multilingue avec des contrôles et validations exercés par les pays utilisateurs.

réaliser des systemès de traduction assistée par ordinateur (TAO)

améliorer les postes de travail de TAO.

assurer la formation nécessaire des utilisateurs potentiels des nouvelles technologies de l'information.

3.5 Base de connaissances

organiser et structurer la sélection et l'acquisition de l'information passée, présente et future à archiver dans une banque de données nationale (savoir faire/matière grise d'un pays) tout en protégeant les caractéristiques de confidentialité de l'information mise en circulation.

définir la structure et le contenu des bases de données.

4. CONCLUSION

Voici très schématiquement les recommandations principales que nous pourrions transmettre à nos autorités compétentes. Il nous parait difficile de préciser un ordre d'urgence dans les recommandations présentées ci-avant. Mais si l'on veut combler les lacunes dans le domaine de la communication de l'information sous toutes ses formes, il est évident que toutes ces recommandations devront être développées et utilisées très rapidement.

Cependant, lorsque l'on examine la liste des recommandations, on peut constater que des actions sont déjà entreprises pour l'ensemble des sujets traités, avec des plans de financement, qui devraient permettre dans la majorité des cas, de finaliser et de commercialiser les projets étudiés par les concepteurs de systèmes.

Alors pour quelles raisons devons nous établir des recommandations ?

1) *Parce que ce sont les utilisateurs qui doivent faire connaître leurs besoins et être associés aux concepteurs, à l'étude et au développement des systèmes dont ils ont besoin pour l'accomplissement de leurs taches.*

2) *Tous les systèmes développés doivent pouvoir s'intégrer dans l'environnement bureautique existant, et surtout, être compatibles entre eux et s'intégrer dans un réseau général d'informations. D'où l'importance d'une coordination des travaux par un organisme officiel.*

3) *Pour convaincre nos autorités de la nécessité de créer un organisme officiel coordonnateur, chargé de définir une politique documentaire et d'en assumer sa mise en application.*

Voici donc très brièvement quelques raisons qui nous poussent à établir des recommandations que nous pourrions adresser à nos autorités et les actions à entreprendre, en veillant plus particulièrement à la formation du personnel qui sera amené à utiliser toutes ces nouvelles technologies.

BRIDGING THE COMMUNICATION GAP

Technical Information Panel Specialists' Meeting
Trondheim, Norway, 5—6 September, 1990

LIST OF PARTICIPANTS

Mr K.E.ANDERSEN	Nasjonal Bibliotekaudelinga i Rana, Postboks 278, N-8601 Mo, Norway
Ms T.AURSOY	The Technical University Library of Norway, N-7034 Trondheim, Norway
Ms I.BANG	The Technical University Library of Norway, N-7034 Trondheim, Norway
Col. F.BARBIERI*	Ministero Della Difesa, Costarmaereo — UCT, Direzione Generale Costruzioni AAAS, Viale dell'Universita 4, 00185 Rome, Italy
Ms M.J.BARRULAS†	Department of Information Studies, University of Sheffield, Sheffield S10 2TN, United Kingdom
Ing. N.BENAMOU†	Societe SELISA, 17 Av. du Parc, 91380 Chilly Mazarin, France
Mr C.J.BIGGER*	Chief Librarian, GEC MARCONI Research Centre, West Hanningfield Road, Great Baddow, Chelmsford, Essex CM2 8HN, United Kingdom
Mr T.BORRESEN	Forsvarets Skole i Etterretniwg og Sikkerhetstjeneste, Oslo Mil Leven, 0018 Oslo 1, Norway
Mr M.BRANDRETH*	Chief, Policy, Planning & Systems, Canada Institute for Scientific and Technical Information, National Research Council of Canada, Ottawa, Ontario K1A OS2, Canada
Lt Col. H.BRAUN*	Dokumentations- und Fachinformationszentrum der Bundeswehr, Friedrich-Ebert-Allee 34, D-5300 Bonn 1, Germany
Mr B.BRINTET	Aerospatiale, Centre d'Infor. Documentaire Commun, B.P. 76, 921252 Suresnes Cedex, France
Mr L.CANIVEZ†	Societe SELISA, 1 Boulevard Arago, Zone Industrielle de Villemilan, 91320 Wissous, France
Ms B.C.CARROLL	President, Information International, P.O. Box 141, Oak Ridge, TN 37831, United States
Mr J.-P.CHATEAU	ONERA, B.P. 72, 92322 Chatillon-Cedex, France
Ing.Gen. F.CHEVALIER*	Directeur, C.E.D.O.C.A.R., 00460 Armees, France
Ms Z.CORREIA†	LNETI, Research Assistant CITI, Azinhaga dos Lameiros, 1699 Lisboa, Codex, Portugal
Dr Ing. A.M.R.CORREIA*	LNETI, Director, CITI, Azinhaga dos Lameiros, 1699 Lisboa, Codex, Portugal
Ms W.DAHL NERBO	The Technical University Library of Norway, N-7034 Trondheim, Norway
Dr A.DEL REY*	I.C.Y.T. (CSIC), Head U.E.I. Information Retrieval, c/ Joaquin Costa 22, 28002 Madrid, Spain
Mr G.DI MARTINO*	CIRA, Relazioni Esterne, Via Boncompagni 93, 00187 Rome, Italy
Mr G.DIKVOLD	Nasjonal Bibliotekavdelinga i Rana, Postboks 278, N-8601 Mo, Norway
Mr A.DRAGSTEN	University Library of Trondheim, Erling Skakkes Gt 47C, 7014 Trondheim, Norway
Prof. E.H.EGEBERG	ISL/Universitetet, N-9000 Tromso, Norway
Mr D.ELAZAR	Israeli Aircraft Industries Ltd., Technical Information Centre — 2416, Ben Gurion International Airport, 70100 — Israel
Mr S.D.ELPHICK	Head of the Library, The Royal Netherlands Naval Academy, Het Nieuwe Diep 8, 1781 AC Den Helder, Netherlands
Mr P.A.EVJEMO	The Technical University Library of Norway, N-7034 Trondheim, Norway
Mr J.FENNER	FhG-INT, Appelsgarten 2, D-5350 Euskirchen, Germany
Mr E.FLOOD	The Technical University Library of Norway, N-7034 Trondheim, Norway
Mr J.J.GEORGET	Aerospatiale (DSSS), MU/DTD, B.P. 96, 78133 Les Mureaux Cedex, France

† Author
* Panel Member

Mrs J.A.GIBSON	Director, US Army TRADOC Analysis Command, ATTN: ATRC-WSR, White Sands Missile Range, NM-88002-5502, United States
Mrs R.A.GJERSVIK*	Director of the Library, The Technical University, Library of Norway, Hogskoleringen 1, N-7034 Trondheim — NTH, Norway
Mr N.GRAM	Riksbibliotektjenesten, Bygdoy alle 21, N-0262 Oslo 2, Norway
Mr B.GUILLOT	SNECMA, TPXD — Chef du Service, Information Documentaire, Etablissement de Villaroche, 77550 Moissy Cramayel, France
Dr M.C.GUTIERREZ*	INTA, Paseo Pintor Rosales 34, 28008 Madrid, Spain
Mr P.A.HAIGH	Head of Collections, Library Document Supply Centre, Boston Spa, Wetherby, Yorkshire LS 23 7BQ, United Kingdom
Mr K.B.HANSEN	SINTEF NORTH-NORWAY, Postboks 250, N-8501 Narvik, Norway
Mr B.HEGSETH	The Technical University Library of Norway, N-7034 Trondheim, Norway
Mr M.HEPWORTH†	Profile Information, P.O. Box 12, Sunbury-on-Thames, Middlesex TW16 7UD, United Kingdom
Mr R.HICKMAN	AGARD Translator, 7 rue Ancelle, 92200 Neuilly-sur-Seine, France
Ms H.HINZ	Copenhagen Business School, Dept. of Computational Linguistics, Dalgas Have 15, DK-2000 Frederiksberg, Denmark
Mr B.HISINGER M.Sc.E.E.*	Senior Consultant/The National Technological Library Denmark, c/o Technical University of Denmark, Anker Engelunds Vej 1, 2800 Lyngby, Denmark
Mr A.HOFF	Trondheim Folkebibliotek, N-7004 Trondheim, Norway
Ir P.J.HOOGENBERK*	Ministerie van Defensie, Hoofd TDCK, Postbox 90701, 2509 LS The Hague, Netherlands
Mr S.JOHANSEN	Research Librarian, Universitetsbiblioteket i Trondheim, N-7004 Trondheim, Norway
Mme D.Jule	Aerospatiale, E/DE/D, 12 rue Beranger, 92320 Chatillon-Bagneux, France
Col. D.KAYA*	Ministry of National Defence (MSB), Chief, Dept. of R&D (ARGE), 06550 Ankara, Turkey
Mrs G.KJELLDAHL	Teknologisk inst. Biblioteket, P.B. 2608, St. Hanshaugen, N-0131 Oslo, Norway
Ms T.KNUTSEN	The Technical University Library of Norway, N-7034 Trondheim, Norway
Ms A.LAMVIK	The Technical University Library of Norway, N-7034 Trondheim, Norway
Mrs R.LANDE	Chief Librarian, Medisinsk Bibliotek, Universitetet i Trondheim, Regionsykehuset, N-7006 Trondheim, Norway
Mr P.LAVAL†	Societe CORA, 93 Av de Fontainebleau, 94270 Le Kremlin Bicetre, France
Mr O.LAVROFF*†	5 rue Anna Jacquin, 92100 Boulogne, France
Ms B.LAWRENCE*†	Administrator, Technical Information Officer, A.I.A.A., 555 West 57th Street, 12th Floor, New York, NY 10019, United States
Mr C.LEBLOND	Dassault Electronique, 55 Quai Marcel Dassault, 92214 Saint Cloud, France
Mr G.-F. LECOQ	CIGREF, 21 Avenue de Messine, 75008 Paris, France
Mr M.E.LESK†	Bell Communications Research, Room 2A-385, 435 South Street, Morristown, NJ 07960, United States
Mme F.LHULLIER*	Chef du Service Documentation, ONERA, 29 Av. de la Division Leclerc, 92320 Chatillon, France
Mr I.LOMHEIM	The Technical University Library of Norway, N-7034 Trondheim, Norway
Professor B.MAEGAARD†	Head of EUROTRA-DK, University of Copenhagen, Njalsgade 80, DK-2300 Copenhagen S., Denmark
Dr G.MAHE	Chef de Departement, CELAR, CCSA/iD, 35170 Bruz, France
Dr J.MARSHALL HUGHES II	Naval Surface Warfare Center, Technical Library Code E 23, Dahlgren, VA 22448-5020, United States
Dipl.Phys. H.MEIER*	Fachinformationszentrum Karlsruhe, D-7514 Eggenstein-Leopoldshafen 2, Germany

† Author
* Panel Member

Mr P.MESNAGER	Chef du Service Information, SEP, B.P. No. 303, 92156 Suresnes, France
Mr K.N.MOLHOLM*	Administrator, Defense Technical Information Ctr., Cameron Station, Alexandria, VA 22304-6145, United States
Mr H.NORDLIE	Editor in Chief, Cappelen, Boks 350 Sentrum, N-0101 Oslo 1, Norway
Major S.PAPADIMITRIOU*	Hellenic Air Force General Staff, B' Branch — B3 Directorate Holargos, TGA 1010, Athens, Greece
Mr B.PAVIOT	GSI-ERLI, 1 Place des Marseillais, 94227 Charenton-le-Pont, France
Ing. P.PELLEGRINI†	Societe CORA, 93 Ave de Fontainebleau, 94270 Le Kremlin Bicetre, France
Colonel T.REDMOND	USAF, Military Committee Studies Division, AGARD, 7 rue Ancelle, 92200 Neuilly-sur-Seine, France
Mr L.ROLLING†	CCE, DG 13, Batiment Jean Monnet, Luxembourg 2920, Luxembourg
Lt Rui A.G.B.ROQUE*	CDIFA, Base de Alfragide, Av. Leite de Vasconcelos, 2700 Amadora, Portugal
Mr P.ROUSSEAUX	Chef du Service Documentation, Dassault, 78 quai Marcel Dassault, 92214 Saint Cloud, France
Ms R.SANNINO	C.I.R.A., Via Maiorise, 81043 Capua (CE), Italy
Mr M.J.SCHRYER*	Director, Directorate of Scientific Information Services, National Defence Headquarters, MGeneral George R.Pearkes Building, Ottawa, Ontario K1A OK2, CANADA
Mrs D.L.SCHRYER	41 Pheasant Run Dr., Nepean Ontario, Canada
Mr R.SEARLE*	Chief Librarian, Royal Aerospace Establishment, Farnborough, Hants GU14 6TD, United Kingdom
Mr H.SELBERG	The Technical University Library of Norway, N-7034 Trondheim, Norway
Mrs E.SKAGEN	Librarian, Norges Tekniske Universitetsbibliotek, N-7034 Trondheim, Norway
Mrs V.M.SKARSTEIN	Library Director, Trondheim Folkebibliotek, N-7004 Trondheim, Norway
Mr R.W.SLANEY	Defence Operational Analysis Estab., Ministry of Defence, Parvis Road, West Byfleet, Surrey KT14 6LY, United Kingdom
Mr O.STAMNES	The Technical University Library of Norway, N-7034 Trondheim, Norway
Cdt. J.STERKEN*	Centrale Bibliotheek MLV, KKE/BLOK 6, Everestraat 1, B-1140 Brussels, Belgium
Professor G.STETTE†	Norwegian Institute of Technology, N-7034 Trondheim, Norway
Mr A.K.STEWART	Information Retrieval and Analysis Amoco Research Center, Warrenville Road and Mill Street, P.O. Box 3011, Naperville, Illinois 60566, United States
Mr R.STORLEER†	The Technical University Library of Norway, Documentation Department, Hogskoleringen 1, N-7034 Trondheim, Norway
Ir A.S.T.TAN*	Information Specialist, National Aerospace Laboratory (NLR), P.O. Box 90502, 1006 BM Amsterdam, Netherlands
Ms N.C.TOROSSIAN	Dassault, 78 quai Marcel Dassault, DGT/Documentation, 92214 Saint Cloud, France
Mrs E.URUNDUL*	Tubitak-Turdok, Tunus Caddessi 33, Kavaklidere-Ankara, Turkey
Mrs A.VICKERY†	Tome Associates, IMO House, 222 Northfield Avenue, London W13 9SJ, United Kingdom
Mr B.VICKERY	Tome Associates, IMO House, 222 Northfield Avenue, London W13 9SJ, United Kingdom
Mr G.C.VIS	FEL-TNO Library, P.O. Box 96864, 2509 JG The Hague, Netherlands
Miss C.WALKER*	Head, Information Services Branch, SHAPE Technical Centre, P.O. Box 174, 2501 CD The Hague, Netherlands
Mr H.WENDT†	Head of Department of New Technologies & Product Development, Springer Verlag, Tiergartenstrasse 17, D-6900 Heidelberg 1, Germany
Mr M.R.C.WILKINSON*	Head, Defence Research Information Centre, Kentigern House, 65 Brown Street, Glasgow G2 8EX, United Kingdom

† Author
* Panel Member

Mr A.YANEZ* Conseiller du Directeur, C.E.D.O.C.A.R., 00460 Armees, France

Mrs ZURN† Springer Verlag GmbH & Co. KG, Tiergartenstrasse 17, D-6900 Heidelberg 1, Germany

† Author
* Panel Member

REPORT DOCUMENTATION PAGE

1. Recipient's Reference	2. Originator's Reference	3. Further Reference	4. Security Classification of Document
	AGARD-CP-487	ISBN 92-835-0604-9	UNCLASSIFIED

5. Originator Advisory Group for Aerospace Research and Development
North Atlantic Treaty Organization
7 rue Ancelle, 92200 Neuilly sur Seine, France

6. Title BRIDGING THE COMMUNICATION GAP

7. Presented at the Technical Information Panel Specialists' Meeting, held at the Nye Sentrum Hotel, Trondheim, Norway, 5th to 6th September 1990.

8. Author(s)/Editor(s)	9. Date
Various	February 1991
10. Author's/Editor's Address	**11. Pages**
Various	128

12. Distribution Statement This document is distributed in accordance with AGARD policies and regulations, which are outlined on the Outside Back Covers of all AGARD publications.

13. Keywords/Descriptors

Documentation
Information centres
Information retrieval
Information systems
Telecommunication
Machine translation
Subject indexing
Publishing
Communications management

14. Abstract

Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Trondheim, Norway, 5th to 6th September 1990.

Topics discussed include optical media, computer-assisted publishing, broader communication bands, natural language processing and machine translation, full text retrieval, intelligent on-line interfaces, full text indexing and retrieval, the problems faced by database publishers, and the training of information specialists. The final paper suggested actions that might be taken by information centre managers, and recommendations they might make to ministries and other official bodies.

AGARD Conference Proceedings No.487 Advisory Group for Aerospace Research and Development, NATO BRIDGING THE COMMUNICATION GAP Published February 1991 128 pages Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Trondheim, Norway, 5th to 6th September 1990. Topics discussed include optical media, computer-assisted publishing, broader communication bands, natural language processing and machine translation, full text retrieval, intelligent on-line interfaces, full text indexing and retrieval, the problems faced by database publishers, P.T.O.	AGARD-CP-487 Documentation Information centres Information retrieval Information systems Telecommunication Machine translation Subject indexing Publishing Communications management	AGARD Conference Proceedings No.487 Advisory Group for Aerospace Research and Development, NATO BRIDGING THE COMMUNICATION GAP Published February 1991 128 pages Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Trondheim, Norway, 5th to 6th September 1990. Topics discussed include optical media, computer-assisted publishing, broader communication bands, natural language processing and machine translation, full text retrieval, intelligent on-line interfaces, full text indexing and retrieval, the problems faced by database publishers, P.T.O.	AGARD-CP-487 Documentation Information centres Information retrieval Information systems Telecommunication Machine translation Subject indexing Publishing Communications management
AGARD Conference Proceedings No.487 Advisory Group for Aerospace Research and Development, NATO BRIDGING THE COMMUNICATION GAP Published February 1991 128 pages Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Trondheim, Norway, 5th to 6th September 1990. Topics discussed include optical media, computer-assisted publishing, broader communication bands, natural language processing and machine translation, full text retrieval, intelligent on-line interfaces, full text indexing and retrieval, the problems faced by database publishers, P.T.O.	AGARD-CP-487 Documentation Information centres Information retrieval Information systems Telecommunication Machine translation Subject indexing Publishing Communications management	AGARD Conference Proceedings No.487 Advisory Group for Aerospace Research and Development, NATO BRIDGING THE COMMUNICATION GAP Published February 1991 128 pages Comprises the papers presented at a Specialists' Meeting held by the Technical Information Panel of AGARD in Trondheim, Norway, 5th to 6th September 1990. Topics discussed include optical media, computer-assisted publishing, broader communication bands, natural language processing and machine translation, full text retrieval, intelligent on-line interfaces, full text indexing and retrieval, the problems faced by database publishers, P.T.O.	AGARD-CP-487 Documentation Information centres Information retrieval Information systems Telecommunication Machine translation Subject indexing Publishing Communications management

and the training of information specialists. The final paper suggested actions that might be taken by information centre managers, and recommendations they might make to ministries and other official bodies. ISBN 92-835-0604-9	and the training of information specialists. The final paper suggested actions that might be taken by information centre managers, and recommendations they might make to ministries and other official bodies. ISBN 92-835-0604-9
and the training of information specialists. The final paper suggested actions that might be taken by information centre managers, and recommendations they might make to ministries and other official bodies. ISBN 92-835-0604-9	and the training of information specialists. The final paper suggested actions that might be taken by information centre managers, and recommendations they might make to ministries and other official bodies. ISBN 92-835-0604-9

AGARD

NATO OTAN

7 RUE ANCELLE · 92200 NEUILLY-SUR-SEINE

FRANCE

Téléphone (1)47.38.57.00 · Télex 610 176
Télécopie (1)47.38.57.99

DIFFUSION DES PUBLICATIONS
AGARD NON CLASSIFIEES

L'AGARD ne détient pas de stocks de ses publications, dans un but de distribution générale à l'adresse ci-dessus. La diffusion initiale des publications de l'AGARD est effectuée auprès des pays membres de cette organisation par l'intermédiaire des Centres Nationaux de Distribution suivants. A l'exception des Etats-Unis, ces centres disposent parfois d'exemplaires additionnels; dans les cas contraire, on peut se procurer ces exemplaires sous forme de microfiches ou de microcopies auprès des Agences de Vente dont la liste suite.

CENTRES DE DIFFUSION NATIONAUX

ALLEMAGNE
Fachinformationszentrum,
Karlsruhe
D-7514 Eggenstein-Leopoldshafen 2

BELGIQUE
Coordonnateur AGARD-VSL
Etat-Major de la Force Aérienne
Quartier Reine Elisabeth
Rue d'Evere, 1140 Bruxelles

CANADA
Directeur du Service des Renseignements Scientifiques
Ministère de la Défense Nationale
Ottawa, Ontario K1A 0K2

DANEMARK
Danish Defence Research Board
Ved Idraetsparken 4
2100 Copenhagen Ø

ESPAGNE
INTA (AGARD Publications)
Pintor Rosales 34
28008 Madrid

ETATS-UNIS
National Aeronautics and Space Administration
Langley Research Center
M/S 180
Hampton, Virginia 23665

FRANCE
O.N.E.R.A. (Direction)
29, Avenue de la Division Leclerc
92320, Châtillon sous Bagneux

GRECE
Hellenic Air Force
Air War College
Scientific and Technical Library
Dekelia Air Force Base
Dekelia, Athens TGA 1010

ISLANDE
Director of Aviation
c/o Flugrad
Reykjavik

ITALIE
Aeronautica Militaire
Ufficio del Delegato Nazionale all'AGARD
3 Piazzale Adenauer
00144 Roma EUR

LUXEMBOURG
Voir Belgique

NORVEGE
Norwegian Defence Research Establishment
Attn: Biblioteket
P.O. Box 25
N-2007 Kjeller

PAYS-BAS
Netherlands Delegation to AGARD
National Aerospace Laboratory NLR
Kluyverweg 1
2629 HS Delft

PORTUGAL
Portuguese National Coordinator to AGARD
Gabinete de Estudos e Programas
CLAFA
Base de Alfragide
Alfragide
2700 Amadora

ROYAUME UNI
Defence Research Information Centre
Kentigern House
65 Brown Street
Glasgow G2 8EX

TURQUIE
Millî Savunma Başkanlığı (MSB)
ARGE Daire Başkanlığı (ARGE)
Ankara

LE CENTRE NATIONAL DE DISTRIBUTION DES ETATS-UNIS (NASA) NE DETIENT PAS DE STOCKS DES PUBLICATIONS AGARD ET LES DEMANDES D'EXEMPLAIRES DOIVENT ETRE ADRESSEES DIRECTEMENT AU SERVICE NATIONAL TECHNIQUE DE L'INFORMATION (NTIS) DONT L'ADRESSE SUIT.

AGENCES DE VENTE

National Technical Information Service
(NTIS)
5285 Port Royal Road
Springfield, Virginia 22161
Etats-Unis

ESA/Information Retrieval Service
European Space Agency
10, rue Mario Nikis
75015 Paris
France

The British Library
Document Supply Division
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
Royaume Uni

Les demandes de microfiches ou de photocopies de documents AGARD (y compris les demandes faites auprès du NTIS) doivent comporter la dénomination AGARD, ainsi que le numéro de série de l'AGARD (par exemple AGARD-AG-315). Des informations analogues, telles que le titre et la date de publication sont souhaitables. Veuiller noter qu'il y a lieu de spécifier AGARD-R-nnn et AGARD-AR-nnn lors de la commande de rapports AGARD et des rapports consultatifs AGARD respectivement. Des références bibliographiques complètes ainsi que des résumés des publications AGARD figurent dans les journaux suivants:

Scientifique and Technical Aerospace Reports (STAR)
publié par la NASA Scientific and Technical
Information Division
NASA Headquarters (NTT)
Washington D.C. 20546
Etats-Unis

Government Reports Announcements and Index (GRA&I)
publié par le National Technical Information Service
Springfield
Virginia 22161
Etats-Unis

(accessible également en mode interactif dans la base de données bibliographiques en ligne du NTIS, et sur CD-ROM)

Imprimé par Specialised Printing Services Limited
40 Chigwell Lane, Loughton, Essex IG10 3TZ